Increasing X-Ray Image Interpretation Competency in Aviation Security with Computer-Based Training

Thesis
Presented to the Faculty of Arts
of
the University of Zurich
for the Degree of Doctor of Philosophy

by

Stefan Michel

of Netstal GL

Accepted in the Fall Semester 2008
on the Recommendation of
Prof. Dr. Wolfgang Marx & Prof. Dr. Martin Kleinmann

ETH Press, Zurich

2008

Contents

1 Summary

The ongoing history of terrorists' pattern in airports and on airplanes shows the importance of scientific research in that field, especially regarding x-ray image interpretation. X-ray technologies provide us with the inside view of luggage and allow to see the content of a bag without opening it. Drastic growth in such technologies has been made over the past few years. The quality of x-ray images is getting better and better, and especially the image resolution has improved substantially. Moreover a screener can use different image enhancement functions (IEFs) which should help screeners to find prohibited threat items in passenger bags more easily. State-of-the-art x-ray machines provide automatic explosive detection algorithms, but they still have high false alarm rates. Therefore, the inspection task cannot be solved by the machine itself; the human operator has always the last decision. This decision is determined by the level of human performance which needs to take into account perceptual and cognitive characteristics. The goal is to achieve a consistent and high performance which can only be achieved when screeners are well trained and familiarized with different kinds of threat objects. The most expensive equipment is of limited value if the screener who operates it is not selected and trained appropriately.

In this dissertation I focus on psychological aspects of a computer-based training system considering the scientific background of perceptual psychology, psychophysics, signal detection theory and human machine interactions.

First, the question arises if so-called image enhancement functions (IEFs) really enhance threat detection performance of screeners. According to manufacturers, a screener should be able to achieve a higher detection

performance using such IEFs compared to the original image. In Chapter 2, I will discuss a study in which I could show that the original image always leads to better detection performance compared to the use of IEFs. These results extenuate the use of IEFs, especially during training: it takes more time to use them and therefore the learning process is weakened. In Chapter 3, automated image difficulty estimations are discussed. These automated image difficulty estimations are very important assumptions for an individually adaptive training system, because for each screener the difficulty of a bag and of a bag-to-threat combination can be calculated and therefore the progress in training can be optimally adapted individually to each screener. One of the main goals in this dissertation is to investigate how to increase screeners' detection performance most effectively with computer based training. An overview of how screeners can be trained and how the effectiveness of such a training system can be proved is discussed in chapter 4. In different studies I could show that individually adaptive computer-based training (CBT) leads very effectively to an increased performance for detecting prohibited items in passenger bags. Moreover, transfer of knowledge from learned to new threat objects could also be proven. This is very important considering that a screener has not the possibility to learn every threat object existing in the world. Furthermore, the benefits and costs of a new technology which allows multi-view images of a passenger bag is empirically investigated and discussed in chapter 5. I could show that for some difficult conditions of bag-to-threat combinations, this technology can help to find threat objects more easily. An outlook of the usefulness of developing a training system for such a technology is contemplated. Finally, recommendations for training level achievement and training duration are discussed in chapter 6. These recommendations are very important for authorities and also for supervisors when guidelines are prepared. All in all, I could show that individually adaptive training is one of the most important factors when screeners have to detect threat items in passenger bags reliably and effectively.

2 Image Enhancement Technology

2.1 Do "Image Enhancement" Functions Really Enhance X-Ray Image Interpretation?

2.1.1 Abstract

State-of-the art x-ray screening systems offer a variety of so-called “image enhancement” functions (IEFs). Examples are color inversion, edge-enhancement, organic only, metal only etc. IEFs are often promoted because they would bring out detail that is obscured or highlight certain features, such as for example organic content. In this study, we investigated the usefulness of IEFs for cabin baggage screening (CBS) and hold baggage screening (HBS) in airport security. The results showed that the standard image provided the best detection performance. Some IEFs impaired detection performance substantially, which was also dependent on threat type (guns, knives, improvised explosive devices, other threat items). Together with previous work (Klock, 2005), these results highlight the importance of systematically studying the usefulness of IEFs in order to optimize human-computer interaction in x-ray screening.

2.1.2 Introduction

In recent years, the importance of baggage x-ray screening at airports has increased dramatically. The image quality of older x-ray screening equipment was sometimes in need of improvement. For example an early version of a coloring algorithm as enhancement function did not serve the purpose of increasing detection performance of threat objects, actually it impaired it. This was due to the occlusion of object parts by the opaque coloring algorithm (Schwaninger, 2005a; Schwaninger, 2005b). But there has been much technological progress in the last years, especially regarding x-ray screening machines, which nowadays provide high image quality and various image enhancement functions (IEFs). The main objective of such functions is to process an image so that the result is more suitable than the original image for a specific application as for example x-ray screening at airports (Gonzalez & Woods, 2002). In x-ray images, the image enhancements might increase the visibility of objects within the bag and remove background noise. State-of-the-art x-ray machines provide many IEFs. The aim of this study is to investigate whether IEFs actually help human operators (screeners) to better detect threat items in x-ray images of passenger bags. Interestingly, reports regarding an evaluation of IEFs have not been publicly available except two recent publications (Klock, 2005; Schwaninger, 2005b). Klock (2005) examined whether IEFs increase screeners' threat detection performance when visually inspecting carry-on bags using a Rapiscan emulator. She found that high penetration, organic stripping and inorganic stripping functions resulted in decreased probability of detection (see below for more information on different IEFs). Crystal clear, black and white, and low penetration resulted in the best performance, while it should be noted that the original color image was not included in the analysis. Klock (2005) also found that these effects are dependent on threat type, i.e. whether guns, knives, improvised explosive devices (IEDs) or other prohibited items had to be detected. Schwaninger (2005b) reported a study on the effects of IEFs for the detection of IEDs in hold baggage. He found that the original image resulted in the best performance, whereas the organic stripping, organic only and luminance negative functions substantially impaired detection of IEDs. The purpose of this study is to extend previous research in order to evaluate the value of different IEFs. In addition, a comparison between IEFs used in cabin baggage screening (CBS) and hold baggage screening (HBS) was of interest.

The nine IEFs examined in this study can be applied to the x-ray images online when working at an aviation security checkpoint. Each pixel in the image format used in these x-ray machines has a material and a luminance value. To show the images on a screen, the pixel values are color coded using red for organic, blue for metallic and green for mixed organic/metallic material. The luminance value defines the luminance of the pixel. Table 1 gives an overview and description of all IEFs used in Experiment 1.

Table 1: Image Enhancement Filters

Grayscale (GR) 	The Grayscale filter removes the material information from the image and shows only the luminance value.
Luminance High (LH) 	In this filter, the luminance of the image is increased by applying a gamma correction (Pratt, 2001) to the luminance value. This allows the screeners to see details in dark areas of x-ray images, but as a consequence the visibility of details in light areas of the images is reduced.
Luminance Low (LL) 	As the opposite of the Luminance High filter, the luminance of the image is decreased. Details in light areas of the image become more visible, dark areas lose the details.

Luminance Negative (LN) 	In the Luminance Negative filter, the luminance of the image is inverted. The material value and therefore the hue of each pixel remains the same.
Metal Only (MO) 	Here, only the metallic parts of the image are shown in color. The organic parts are transformed to light gray with low contrast. The organic parts of the mixed organic/metallic pixels are removed as well, giving them a blue color similar to the all-metallic parts. The motivation for this filter is to allow the screeners to concentrate on the metallic objects perhaps leading to less search time for such objects.
Metal Stripping (MS) 	The Metal Stripping filter removes the metal from the image. Metallic parts are transformed to light gray and from the mixed organic/metallic pixels the metallic part is removed. As some mixed organic-metallic parts originate from metallic objects laying upon organic objects, this removal of metal sometimes shows the complete organic object without potentially distracting metallic parts.
Organic Only (OO) 	The Organic Only filter shows the organic parts of the image in color, while the metallic pixels are set to gray. The mixed organic/metallic pixels are assigned to the metallic or organic parts depending on the proportion of metallic and organic material. The difference to the Metallic Stripping

	filter is that less of the image remains visible and that the remaining mixed organic/metallic pixels are still green.
Original (OR) 	Original (OR) refers to the unaltered images as produced by the x-ray screening machine without applying any image enhancement filter.
Organic Stripping (OS) 	As the opposite to the Organic Only filter, the metallic parts of the image remain colored and the organic parts are shown in light gray with low contrast. The resulting image is similar to the Metal Only image, except that in this filter the mixed organic/metallic pixels are still green.
Super Enhancement (SE) 	The Super Enhancement filter adaptively adjusts the contrast of the image. Similar to a Local Histogram Equalization (Gonzalez & Woods, 2002) or an Adaptive Contrast Enhancement (Stark, 2000), the luminance of each pixel is adjusted to the luminance of its surrounding pixels. In the resulting image, each area has a medium average luminance.

2.1.3 Experiment 1

Experiment 1 was conducted to evaluate IEFs available in conventional cabin baggage screening (CBS).

2.1.3.1 Participants

A total of 443 airport security screeners of the CBS at a European airport participated in this study. All had on-the-job experience of at least 6 months. A between-subjects design was used to compare the effect of the IEFs on detection performance with each other. To this end, participants were randomly assigned to one of nine experimental groups, one group for each of the nine IEFs specified in Table 1. The control group was used for testing detection performance when images were displayed using the Original (OR) image type. The assignment of participants to groups was conducted so that the distribution of gender, age, and days on job were equal across groups. The ten groups showed an equal average of detection performance A', which was calculated using data of a separate test conducted prior to this study. The experimental groups varied in size between 37 (Luminance Negative filter) and 66 screeners (Grayscale filter); the control group consisted of 39 screeners. The difference in the group sizes is due to missing values (i.e. incomplete tests) for several screeners who originally were assigned to the study.

2.1.3.2 Method and Procedure

The X-Ray Competency Assessment Test (X-Ray CAT) was used in Experiment 1. This computer-based test contains 256 x-ray images of real passenger carry-on bags. Half of these images contain one prohibited item. The prohibited items have been selected by police experts to be representative for the variety of different threat types. The test contains 32 x-ray images of passenger bags with guns, 32 images with knives, 32 images with improvised explosive devices (IEDs), and 32 images with other prohibited items. For further details on the X-Ray CAT, see Koller and Schwaninger (2006). In order to create the stimuli for Experiment 1, the nine IEFs explained in Table 1 above were applied to the x-ray images. The participants' task is to visually inspect the images and to judge whether they are OK (contain no prohibited item) or NOT OK (contain a prohibited item). In this study, images disappeared after 10 seconds. The experiment consisted of two blocks. In block 1, each of the 9 experimental groups was tested with only one IEF and the control group was tested with the Original image (OR). The purpose of block 2 was to confirm that the participant groups are equivalent regarding their x-ray image interpretation competency. In block 2, all participants were tested again using the same bags as in block 1 but images were displayed in the OR format (see Table 1).

2.1.4 Results and Discussion

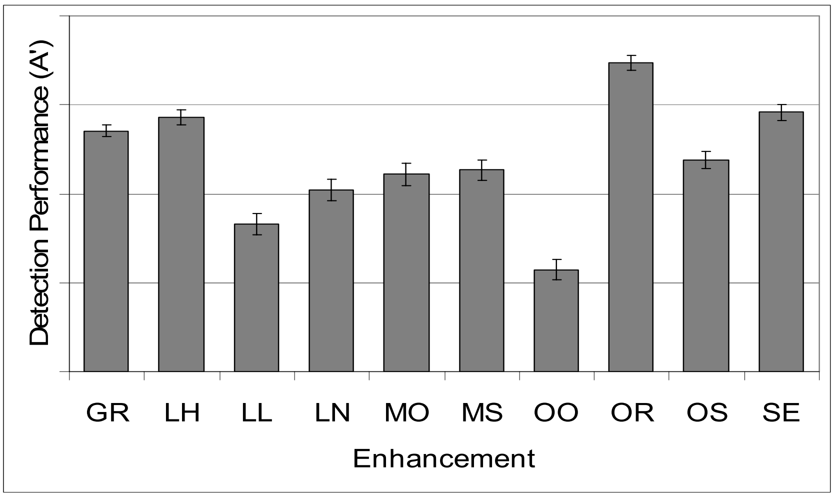

Figure 1: Detection performance Experiment 1, block 1, pooled across threat categories. IEFs were tested between participant groups: GR = Grayscale, LH = Luminance High, LL = Luminance Low, LN = Luminance Negative, MO = Metal Only, MS = Metal Stripping, OO = Organic Only, OS = Organic Stripping, SE = Super Enhancement, OR = Original (control group).

Detection performance was measured using A', a measure derived from hit and false alarm rates (Pollack & Norman, 1964; see Hofer & Schwaninger, 2004 for x-ray image interpretation competency). The hit rate refers to the proportion of all images containing a prohibited item that have been judged as NOT OK. The false alarm rate refers to the proportion of NOT OK judgments for harmless bags. A' scores were calculated for each block separately. Figure 1 shows means and standard errors of A' scores of block

1 broken up by IEF and pooled across threat categories, including the results of the control group (OR). The results in Figure 1 suggest that the OR image type results in the best performance, while some IEFs result in substantial impairment of detection performance. Note that due to security reasons, A' scores are not shown in the figures. To estimate effect sizes we employ effect size analysis and interpret the results based on Cohen (1988).

An analysis of variance (ANOVA) with the between-participant factor IEF was carried out on individual A' scores averaged per screener across threat category. There was a main effect of IEF with a large effect size of $\eta^2 = .46$, $F(9, 433) = 41.67$, $p < .001$.

Figure 2 shows means and standard errors of A' scores of block 1 broken up by IEF and threat category. For all four threat categories, the OR image type resulted in the best performance. Again, some IEFs impaired detection performance substantially. Moreover, the results in Figure 2 suggest that the effects of IEFs on performance vary between threat categories. These results were confirmed by a separate ANOVA using individual A' scores calculated for each of the four threat categories (guns, knives, IEDs, other prohibited items). The ANOVA with the between-participant factor IEF and the within-participant factor threat category gave a large main effect of IEF with an effect size of $\eta^2 = .48$, $F(9, 433) = 43.66$, $p < .001$. There was also a large main effect of threat type with an effect size of $\eta^2 = .30$, $F(3, 1299) = 180.84$, $p < .001$. And there was also a large interaction between threat category and IEF with $\eta^2 = .32$, $F(27, 1299) = 22.91$, $p < .001$.

The same A’ scores were subjected to one-way ANOVAs that were conducted separately for each threat category. There was a large main effect of IEF for all threat categories. For guns, there was an effect size of $\eta2 = .64$, $F(9, 433) = 86.09$, $p < .001$, for IEDs $\eta2 = .32$, $F(9, 433) = 22.38$, $p < .001$, for knives $\eta2 = .32$, $F(9, 433) = 23.10$, $p < 001$, and for other prohibited items $\eta2 = .43$, $F(9, 433) = 36.27$, $p < .001$.

In short, the OR image type resulted in the best performance for all threat categories. Moreover, some IEFs resulted in a substantial impairment which clearly depended on threat category. This interaction would be predicted if one takes into account that color information in x-ray images represents different materials and that different prohibited items vary in their material composition. For example, the Metal Only (MO) filter removes organic material from the x-ray

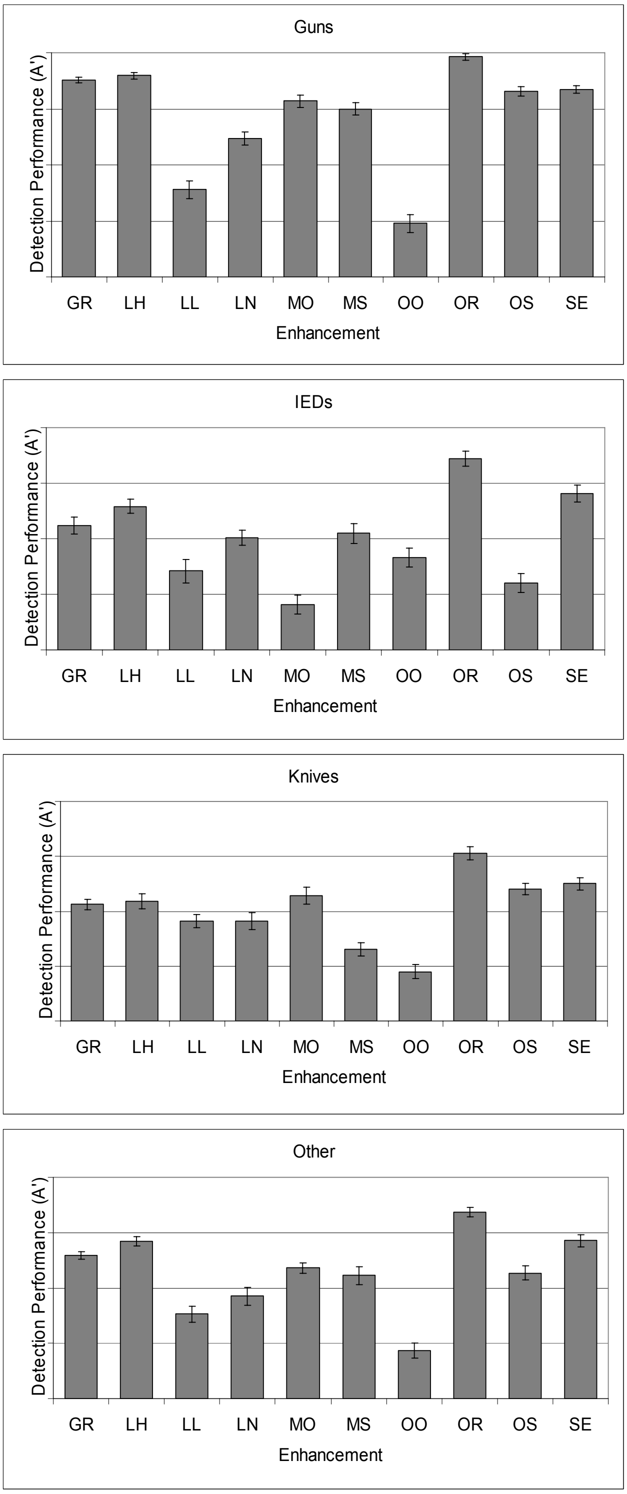

Figure 2: Detection performance in Experiment 1, block 1, broken up by threat category.
GR = Grayscale, LH = Luminance High, LL = Lumi-nance Low, LN = Luminance Negative, MO = Metal Only, MS = Metal Stripping, OO = Organic Only, OR = Original, OS = Organic Stripping, SE = Super Enhancement.

image (see also Table 1). Since guns and knives usually consist of metallic material, their pixels in the filtered x-ray image remain largely unaffected when the MO filter is used. However, explosive material of IEDs is organic, thus it is not surprising that the MO filter results in a large impairment of IED detection (see Figure 2). A similar explanation applies to the effect of the Organic Stripping (OS) filter. When this filter is applied, all metallic parts of the image remain colored and the organic parts are shown in light gray with low contrast. The resulting image is similar to the MO image, except that for this filter the mixed organic/metallic pixels are still green. Since the Metal Stripping (MS) filter removes metallic information from the image, this IEF results in a substantial impairment of the detection of guns and knives, which usually contain much metal. Because organic explosive material in IEDs remains visible when the MS filter is used, IED detection is not affected substantially. The results in Figure 2 also indicate that the MS filter might be a better option than the Organic Only (OO) filter. As explained in Table 1, the MS filter includes information about organic material hidden behind metallic parts, whereas the OO filter simply removes these parts from the image. A comparison between the original image (OR) and the grayscale version gives some indications on the relevance of color information. The removal of the color-coded material information by the Grayscale filter (GR) does impair threat detection, while this effect is less pronounced for the detection of guns. Apparently, the luminance information seems to be more important than the material information. When inserting a threat object into a bag, the part of the bag with the object inside normally becomes darker than its surrounding. This is particularly the case for guns which contain much metallic material. Note however, that the removal of material information can conceal objects with the same luminance but different material than its surrounding. A similar problem appears when using the Super Enhancement (SE) filter. For this IEF, the material information remains the same, but the luminance contrast is slightly reduced which has a negative influence on detection performance. The Luminance High (LH) filter allows better threat detection than the Luminance Low (LL) filter. With the LL filter, most objects inside the bag have a luminance close to black, which generally reduces the differentiation of these objects. When using the Luminance Negative (LN) filter, material and luminance information remain in the image, but the luminance is inverted. The impairment of threat detection when using this IEF shows that screeners perform better with a dark object on a light background than if the luminance is inverted.

The results reported so far refer to block 1. As explained in the method section above, all participants conducted the X-Ray CAT again in block 2 using the original image type (OR). This was conducted to confirm post-hoc that the different participant groups are equivalent in terms of their x-ray image interpretation competency. This a prerequisite for the interpretation of the results reported above involving ANOVAs with IEF as between-participant factor. Separate ANOVAs of the data from block 2 confirmed that the 9 experimental groups and the control group were equivalent. Individual A' scores were calculated for each screener based on all trials of block 2. These data were subjected to a one-way ANOVA with participant group as between-participant factor. All groups were equivalent, since there was no effect of group, $\eta^2 = .02$, $F(9, 433) = 1.08$, $p = .38$. Individual A' scores were calculated also for each threat category separately and this data were then analysed using an ANOVA with participant group as between-participant factor and threat category as within-participant factor.

Again, the results show that the participant groups were equivalent in terms of their x-ray image interpretation competency, since there was no main effect of participant group, and no interaction between participant group and threat category, $\eta^2 = .02$, $F(9, 433) = 1.04$, $p = .41$, and $\eta^2 = .02$, $F(27, 433) = 0.86$, $p = .63$, respectively.

2.1.5 Experiment 2

In hold baggage screening (HBS) x-ray images feature slightly different colors. Figure 3 shows examples of the stimuli used in Experiment 2. As explained in the introduction, screeners mainly search for IEDs, as other threat objects like for example knives do not pose a threat to the aircraft and passengers when placed in hold baggage. HBS screeners are often also more experienced screeners as it was the case in this participant sample. The main aims of Experiment 2 were to examine whether similar results are found in HBS regarding the effect of IEFs despite the operational and training differences between HBS and CBS.

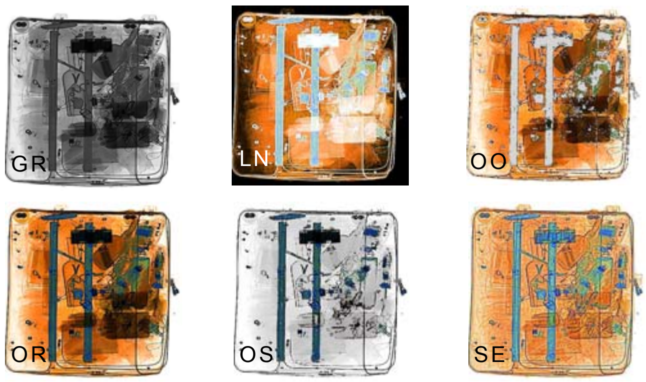

Figure 3: IEFs for HBS as used in Experiment 2. From top left to bottom right: GR, LN, OO, OR, OS, SE (see Table 1).

2.1.5.1 Participants

Data of 83 aviation security screeners of the HBS of the same European airport was analyzed. As in Experiment 1, a between-subjects design was used to compare the effect of the IEFs. Due to the smaller sample size only 5 IEFs and the OR image could be tested. The 83 HBS screeners were randomly assigned to one of five experimental groups (GR, LN, OO, OR, OS, SE filters) or the control group (OR filter). The assignment of participants to groups was conducted so that the distribution of gender, age, and days on job was equal across groups. The six groups showed an equal average of detection performance A', which was calculated using data of a separate test conducted prior to this study. The number of screeners in each experimental group were between 10 (GR) and 17 (OO); the control group (OR) consisted of 15 screeners. As in Experiment 1, the difference in the group sizes is due to missing values (i.e. incomplete tests) for several screeners.

2.1.5.2 Method and Procedure

The Bomb Detection Test (BDT) was used in this study. This computer-based test contains 200 x-ray images of real hold baggage, whereas 100 images contain an IED. The IEDs were created by police experts. Participants were instructed to decide for each x-ray image whether it is OK (does not contain an IED) or NOT OK (contains an IED). Images disappeared after 10 seconds. As in Experiment 1, there were two blocks. In block 1, each of the 5 experimental groups was tested with their respective IEF. In block 2 all participants were then tested again using the same images but using the Original (OR) image function. The control group conducted the test twice using the OR image type in block 1 and block 2. As in Experiment 1, the purpose of block 2 was to confirm the comparability of the groups post hoc.

2.1.6 Results and Discussion

Analyses were similar to Experiment 1 but there was only one threat category, i.e. IEDs. Figure 4 shows means and standard errors of A' scores broken up by image enhancement function. As mentioned above, A' scores are not shown in the figure for security reasons. Effect sizes are calculated using effect size analysis and they are interpreted based on Cohen (1988). A one-way ANOVA with IEF as between-participant factor revealed a large main effect of IEF with an effect size of $\eta^2 = .26$, $F(5, 77) = 5.29$, $p < .001$. As in Experiment 1, the original image (OR) resulted in the best performance. Consistent with the results found in Experiment 1, we found in Experiment 2 that the Organic Stripping (OS) and Luminance Negative (LN) functions resulted in a substantial impairment of detection performance for IEDs.

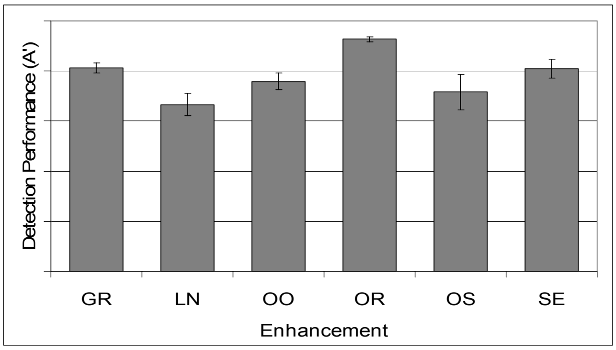

Figure 4: Detection performance Experiment 2, block 1. GR = Grayscale, LN = Luminance Negative, OO = Organic Only, OR = Original, OS = Organic Stripping, SE = Super Enhancement.

All participants conducted the test again in block 2 using the original image type (OR). The aim was to confirm post-hoc that the different participant groups are equivalent in terms of their x-ray image interpretation competency. To this end, individual A' scores from block 2 were subjected to a one-way ANOVA with participant group as between-participant factor. There was no main effect of group, $\eta^2 = .05$, $F(5, 77) = 0.75$, $p = .59$, confirming that the six groups are equivalent regarding their x-ray image interpretation competency.

2.1.7 General Discussion

The aim of this study was to investigate the effect of image enhancement functions (IEFs) on x-ray detection performance of airport security screeners. Experiment 1 was conducted with cabin baggage and Experiment 2 with hold baggage. In both experiments the original image (OR) resulted in the best performance. One interpretation could be that for this manufacturer the default image is indeed the best image. However, since the OR image is the default image on the tested x-ray machine and since screeners received more training with OR images, further research is needed to clarify whether the benefit of the OR image type is due to expertise and training or whether it truly reflects better image quality. In both experiments, it was also found that some IEFs resulted in substantial impairments of detection performance. This general result is consistent with previous reports (Klock, 2005; Schwaninger, 2005b). The IEF effects are dependent on threat category; most likely due to differences in material properties of the different threat categories. For example, guns contain more metal than IEDs. Removing metallic content (MS function) therefore results in a larger impairment of detection performance for guns than for IEDs.

The main conclusions of this study are that user testing is crucial before implementing such filters into a system. Moreover, training when and how to use each of the filters is crucial to make effective use of them. We are conducting a set of additional experiments to further investigate the value of

IEFs. For example, it could be that although on average certain IEFs impair detection, they could still be useful for detecting certain threat objects under certain conditions. Moreover, we are currently looking at CBT data where screeners have the possibility to choose a filter and to switch between filters. This allows investigating whether perhaps a certain combination and sequence of IEFs is useful for certain threat types and images. In addition, we are trying to clarify, whether IEFs actually do not improve detection of prohibited items or if however, when used according to individual preferences and to specific features of the image, they can improve the ability to locate targets. Finally, we also have implemented IEFs in a CBT system (X-Ray Tutor) to investigate potentially supporting effects that only can become manifest through training and familiarization.

2.1.8 Acknowledgments

This research was financially supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403). Many thanks to Zurich State Police, Airport Division, for their help in creating the stimuli and the good collaboration for conducting parts of the study.

2.1.9 References

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. New York: Erlbaum, Hillsdale.

Gonzalez, R.C., & Woods, R.E. (2002). *Digital Image Processing (2nd ed.)*. Upper Saddle River, NJ: Prentice Hall.

Hofer, F., & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *IEEE ICCST Proceedings*, 303-308.

Klock, B.A. (2005). Test and evaluation report for X-ray detection of threats using different X-ray functions. *IEEE ICCST Proceedings*, 182-184.

Koller, S., & Schwaninger, A. (2006). Assessing X-ray image interpretation competency of airport security screeners. *Proceedings of the 2nd International Conference on Research in Air Transportation*, ICRAT 2006, Belgrade, Serbia and Montenegro, June 24-28, 2006, 399-402.

Pollack, I. & Norman, D.A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, 1, 125-126.

Pratt, W. K. (2001). *Digital Image Processing*: PIKS Inside, 3rd Edition. John Wiley & Sons, Inc., New York, NY.

Schwaninger, A. (2005a). Objekterkennung und Signaldetektion: Anwendungen in der Praxis. In B. Kersten & M. Groner (Hrsg.), *Praxisfelder der Wahrnehmungspsychologie*. Bern: Huber.

Schwaninger, A. (2005b). X-ray Imagery: enhancing the value of the pixels. Aviation Security International, Oct 2005, 16-21.

Stark, J.A. (2000). Adaptive image contrast enhancement using generalizations of histogram equalization. *IEEE Transactions on Image Processing*, 9, 5, 889 - 896.

3 Automated Image Difficulty Estimation

3.1 Towards a Model for Estimating Image Difficulty in X-Ray Screening

3.1.1 Abstract

In this study we developed a first computational model for estimating image difficulty of x-ray images of passenger bags. Based on Schwaninger (2003) three image-based factors are proposed as predictors of image difficulty: view difficulty of the threat item, superposition by other objects, and bag complexity (i.e. clutter and transparency of the bag). First, these factors were validated using detection experiments. We then developed computer-based algorithms to estimate the image-based factors automatically. Finally, we could show that our computational model can better explain human performance than human ratings of the image-based factors.

3.1.2 Introduction

The relevance of aviation security has increased dramatically in the last years. One of the most important tasks is the visual inspection of passenger bags using x-ray machines. In this study we investigated the role of image-based factors on human detection of prohibited items in x-ray images. Schwaninger (2003) has proposed in that the following image-based factors influence how difficult it is to detect a threat item in x-ray images: view difficulty of the threat item, superposition by other objects, and bag complexity. This was validated in a study conducted by Schwaninger, Hardmeier & Hofer (2004). In Experiment 1, we replicated these results in order to provide converging evidence for the validity of the assumption of different image-based factors. In Experiment 2, the same x-ray images were rated by human participants for view difficulty, superposition, bag complexity (clutter and transparency), and general difficulty. These human ratings were then correlated with detection performance obtained in the first experiment. In Experiment 3, we developed computer-based algorithms to estimate the image-based factors automatically. These estimates were correlated with human ratings of the same image-based factors (obtained in Experiment 2). Using multiple linear regression analysis, we examined in Experiment 4 whether our computer-based estimates were able to predict human performance from Experiment 1 as good as human ratings from Experiment 2 on the same image-based factors could do so.

3.1.3 Experiment 1

The main aim of Experiment 1 was to replicate the results of Schwaninger, Hardmeier & Hofer (2004), in which it was shown that view difficulty, superposition, and bag complexity influence detection performance substantially.

3.1.4 Method and Procedure

3.1.4.1 Participants

Twelve undergraduates of the University of Zurich participated in this study (5 males, 7 females). None of them had participated in a study with x-ray images before.

3.1.4.2 Procedure

The Object Recognition Test (ORT) was used to analyze the influence of the three image-based factors view difficulty, superposition and bag complexity on human detection performance (for details see Schwaninger, Hardmeier & Hofer (2004) and Hardmeier, Hofer & Schwaninger (2005). X-ray images of passenger bags were shown 4 seconds each. Participants had to decide whether a bag is OK (no threat item present) or NOT OK (threat item present). Using a slider control, participants indicated on a 90 point rating scale how sure they were in their decision (confidence ratings). There were a total of 256 test trials: 16 (8 guns, 8 knives) x 2 (easy vs. difficult view) x 2 (low vs. high superposition) x 2 (low vs. high bag complexity) x 2 (threat bag vs. harmless bag). No feedback was given on test trials. Prior to the test trials, 8 practice trials were presented followed by a presentation of the threat items. The 8 guns were shown for 10 seconds followed by a 10 second screen with the 8 knives. Half of the items were shown in easy view, the other half in difficult view (for further details see Schwaninger, Hardmeier & Hofer, 2004; Hardmeier, Hofer & Schwaninger, 2005).

3.1.4.3 Statistical Analysis

In the study conducted by Schwaninger, Hardmeier & Hofer (2004) detection was measured using A' (for details on this and other detection measures see Green & Sweets, 1996; Hofer & Schwaninger, 2004. In this study, we were interested in developing a computational model to explain detection performance of threats in x-ray images. In Experiment 1 we calculated hit rates for each participant by averaging across threat images. Individual hit rates were subjected to a three-way analysis of variance (ANOVA) with view difficulty, superposition and bag complexity as within-participant factors.

3.1.5 Results and Discussion

The main effects are illustrated in Figure 5. All were highly significant with large effect sizes ($\eta 2$ values). View difficulty: $\eta^2 = .95$, $F(1,11) = 211.2$, $p < .001$; superposition: $\eta^2 = .49$, $F(1,11) = 10.5$, $p < .01$; bag complexity: η^2 =

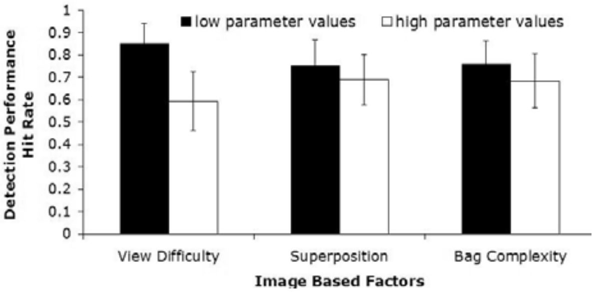

Figure 5: Illustration of main effects of view difficulty, superposition and bag complexity on hit rates for guns.

.59, $F(1,11) = 16.0$, $p < .01$. This replicates earlier findings in which large main effects of view, superposition and bag complexity were found for A' scores (Schwaninger et al., 2004).

Only one significant interaction was found: Bag complexity * view difficulty: $\eta^2 = .77$, $F(1,11) = 36.4$, $p < .001$. All other interactions were not significant. This is consistent with the assumption of three relatively independent factors (whereas only view difficulty and bag complexity might interact).

3.1.6 Experiment 2

3.1.6.1 Introduction

The main aim of Experiment 2 was to investigate whether human ratings of view difficulty, superposition and bag complexity are correlated with human performance measured in Experiment 1.

3.1.7 Method and Procedure

3.1.7.1 Participants

The same participants of Experiment 1 took part in Experiment 2 (with a delay of about one week).

3.1.7.2 Procedure

The same x-ray images as in Experiment 1 were used. The participant's task was to rate view difficulty and superposition of the threat items (threat bags only), and clutter, transparency and general image difficulty (threat and non-threat bags). Ratings were given using a graphical slider control (from very low = 0 to very high = 50). Prior to the ratings, 8 practice trials were presented.

3.1.7.3 Statistical Analysis

Hit rates per x-ray image were calculated by averaging performance data from Experiment 1 across participants. These hit rates were then correlated with x-ray image ratings on threat bags from Experiment 2 (per image, averaged across participants).

3.1.8 Results and Discussion

Pearson correlations showed that ratings of view difficulty and superposition were significantly correlated with hit rate, $r(64) = -.521$, $p < .001$, and $r(64) = -.522$, $p < .001$, respectively. The other correlations did not reach statistical significance: Hit rate and clutter, $r(64) = -.17$, $p = .19$; hit rate and transparency, $r(64) = .08$, $p = .56$. These results could suggest that both clutter and transparency are not relevant for the detection of the threat items used in this study (only guns, see introduction), or that the participants could not reliably estimate the degree of clutter and transparency. We are currently conducting further research to investigate these possibilities.

3.1.9 Experiment 3

3.1.10 Introduction

In Experiment 3, computer-based estimates for image-based factors were developed. They were compared to human ratings from Experiment 2 in order to determine their perceptual plausibility. The following table shows the abbreviations for all independent variables.

Independent Variables	Computer-based Estimates	Rating Estimates
View Difficulty	VD_C	VD_R
Superposition	SP_C	SP_R
Clutter	CL_C	CL_R
Transparency	TR_C	TR_R

Table 2: Abbreviations used in this article. Indices C and R represent computer-based and human rating estimates, respectively.

3.1.11 Method and Procedure

3.1.11.1 Computer based estimates

Computer-based estimates were developed for view difficulty, superposition, and bag complexity (i.e. clutter and transparency).

3.1.11.2 View Difficulty

View difficulty VD was calculated by averaging hit rates ($pHit_i$) across different threat images displaying the same threat item view. In the ORT, each threat item is displayed 4 times from the same viewpoint (see section 3.1.4.2.). The detection performance of the item in question ($pHit_j$) was excluded from this average detection performance. This was done in order to avoid a circular argument in the statistical model by partial inclusion of a predictor into the criterion variable. Therefore, the n in the view difficulty formula equals 4, but the average was calculated over the remaining three (n-1) images displaying the same threat item view.

Formula 1

$$\text{View Difficulty} \quad VD_j = \frac{\left(\left(\sum_{i=1}^{n} pHit_i\right) - pHit_j\right)}{n-1}$$

3.1.11.3 Superposition

The computer-based estimate of superposition is based on the Euclidian distance between the grayscale pixel intensities of the bag with the threat item (ISN) and the bag without it (IN). The following formula was used:

Formula 2
$$Superposition \quad SP = \sqrt{\sum (I_{SN}(x,y) - I_N(x,y))^2}$$

3.1.11.4 Clutter

Clutter (CL) should represent the amount of disarrangement in the bag. In our approach it was estimated based on the amount of high pass frequency information:

Formula 3
$$Clutter\ CL = \sum (I_N(x,y) \otimes F^{-1}(HP(f_x, f_y)))$$
$$HP(f_x, f_y) = 1 - \frac{1}{1 + \left(\frac{\sqrt{(f_x{}^2 + f_y{}^2)}}{f}\right)^d}$$

This convolution ($\otimes$) is equivalent to a high-pass Butterworth filter application in the Fourier-space (F^{-1}: inverse Fourier transform), where f_x and f_y are the frequency components, f is the cut-off frequency and d the fall off.

3.1.11.5 Transparency

Metallic content is more difficult to penetrate by x-ray than organic material, which therefore appears more “transparent” or less opaque in the x-ray image. Transparency was estimated based on the number of pixels in the darkest quarter (< 65) of the pixel intensity range (0 to 255), relative to the bags overall size (areas with pixel intensities ≠ 255).

Formula 4
$$Transparency \quad TR = \frac{\sum_{x,y} (I_N(x,y) < 65)}{\sum_{x,y} (I_N(x,y) \neq 255)}$$

3.1.11.6 Statistical Analysis

To examine the perceptual plausibility of the computer-based estimates we calculated their correlations with the corresponding human ratings from Experiment 2.

3.1.12 Results and Discussion

As can be seen on the diagonal of the correlation matrix of Table 3, all correlations between computer-based estimates and human ratings were highly significant (except for clutter). This shows that at least three of the four of our computer-based estimates of image-based factors are perceptually plausible. The high correlation between computer-based estimates of transparency and human ratings for clutter could indicate that our participants had problems in distinguishing between clutter and transparency. This is consistent with the high correlation between human ratings of clutter and transparency, r(64) = -.79, p < .001.

Table 3: Correlations between computer-based estimates and human ratings.

	VD_R	SP_R	CL_R	TR_R
VD_C	-.61**	-.32**	-.06	-.00
SP_C	-.22	-.44**	-.28*	.15
CL_C	-.04	.12	.15	-.10
TR_C	-.03	.32**	.67**	-.62**

*p<.05. **p<.01

3.1.13 Experiment 4

3.1.14 Introduction

The aim of this experiment was to examine how well our computer-based estimates can explain human performance.

3.1.15 Method and Procedure

Multiple linear regression analysis was used to test how well our computer-based estimates of image-based factors can explain human performance measured in Experiment 1. The human ratings from Experiment 2 were used for benchmarking. More specifically, we tested whether our computer-based estimates of image-based factors achieve a better prediction of human performance than human ratings of the same image-based factors.

3.1.15.1 Statistical Analysis

The two equations below show the two multiple linear regression models using computer-based estimates (C indices) and human ratings (R indices) of image-based factors. The abbreviation DP represents detection performance (hit rate per image averaged across participants), which is the dependent variable. The two models were compared in terms of their goodness-of-fit measures, their regression coefficient's significances, and – most importantly – the percentage of variance in the dependent variable the models were able to explain.

$$DP = b_0 + b_1 VD_C + b_2 SP_C + b_3 CL_C + b_4 TR_C + R$$

$$DP = b_0 + b_1 VD_R + b_2 SP_R + b_3 CL_R + b_4 TR_R + R$$

3.1.16 Results and Discussion

Note that the scales of the computer-based estimates and the rated image based factors have opposite signs. Therefore, the beta-weights in predicting the dependent variable (hit rate per image) have opposite signs in the computational and the rating models.

3.1.17 Computational Model

The computational model correlates with human performance with r = .76 (Figure 6). As shown at the bottom of Table 4, our model using computer-based estimates is able to explain 55 % of the variance of the hit rate (adjusted R^2).

Interestingly, view difficulty and superposition explain most of the variance of the hit rate. In fact, only their beta weights are significant (Table 4).

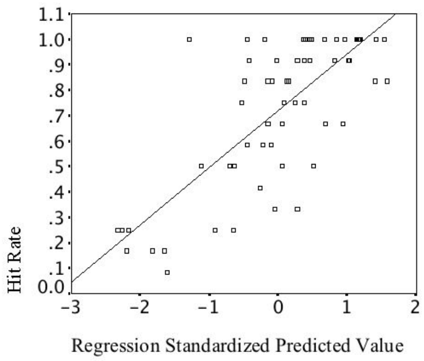

Figure 6: Correlation between predicted and observed performance using the computational model.

Table 4: Summary of regression analysis using computer-based estimates of image-based factors for predicting hit rates.

Variable	B	SE B	β
VD_C	0.78	0.10	.68**
SP_C	0.02	0.01	.23*
CL_C	-0.00	0.00	-.01
TR_C	-0.08	0.45	-.02

R^2 = .581, R^2(adj) = .553, $F(4,59)$ = 20.455, p < 0.001; *p < .05. **p <.01.

3.1.17.1 Human Ratings Model

The model based on human ratings correlates with human performance with r = .70 (Figure 7).

Table 5: Summary of regression analysis using human ratings of image-based factors for predicting hit rates.

Variable	B	SE B	β
VD_R	-0.01	0.00	-.46**
SP_R	-0.02	0.00	-.48**
CL_R	-0.00	0.00	-.05
TR_R	-0.01	0.01	-.11

R^2 = .485, R^2(adj) = .452, $F(4,59)$ = 14.004, p < 0.001; *p < .05. **p < .01.

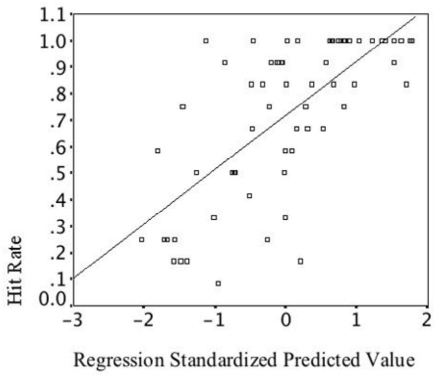

Figure 7: Correlation between predicted and observed performance using the human ratings model.

As can be seen at the bottom of Table 5, the human ratings were able to explain 45 % of the variance of the hit rate (adjusted R^2). This means, that our computational model could explain human performance better than a model based on human ratings. Interestingly, for both models, view difficulty and superposition explained most of the variance of the hit rate and the beta weights of clutter and transparency were not significant.

3.1.18 General Discussion

This study provided converging evidence for the view that detection performance in x-ray screening depends on view difficulty, superposition and bag complexity (Schwaninger, 2003). The results of Experiment 1 showed large main effects of these image-based factors on human detection performance, which is highly consistent with earlier findings by Schwaninger (2003). Human ratings (Experiment 2) and computer-based estimates (Experiment 3) were significantly correlated for view difficulty and superposition. Using multiple regression it was shown in Experiment 4 that our computational model could explain human performance (hit rate) better than a model based on human ratings. Interestingly, for both models, view difficulty and superposition explained most of the variance of the hit rate. In contrast, bag complexity (clutter and transparency) was a weak predictor for both, the model based on human ratings, as well as the computational model. As explained in the introduction, only results from guns are presented in this study. We are currently conducting a series of experiments using different threat types and computer-based estimates in order to extend the computational model presented in this paper and to further investigate the role of bag complexity.

3.1.19 Acknowledgment

We are thankful to Zurich State Police, Airport Divison and Zurich Airport Unique for supporting this study.

3.1.20 References

Schwaninger, A. (2003). Evaluation and selection of airport security screeners. *AIRPORT, 02/2003*, 14-15*.*

Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. *IEEE ICCST Proceedings, 38,* 258-264.

Hardmeier, D., Hofer, F., & Schwaninger, A. (2005). The x-ray object recognition test (x-ray ort) – a reliable and valid instrument for measuring visual abilities needed in x-ray screening. *IEEE ICCST Proceedings*, 39, 189-192.

Green, D.M. & Sweets, J.A. (1996). *Signal detection theory and psychophysics,* New York: Wiley.

Hofer, F. & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *IEEE ICCST Proceedings, 38,* 303-308.

4 How to Increase X-ray Image Interpretation with Computer-Based Training

4.1 Investigating Training and Transfer Effects Resulting from Recurrent CBT of X-Ray Image Interpretation

4.1.1 Abstract

The importance of airport security has increased dramatically in the last years. Large investments into x-ray screening technology have been made in order to cope with the changed terrorist threat situation. However, the most expensive equipment is of limited value if the humans who operate it are not trained well enough to detect threat objects in x-ray images of passenger bags quickly and reliably. In this study we investigated whether adaptive computer based training (CBT) can be used to increase x-ray image interpretation competency of airport security screeners. To this end, we tested screeners before and after six months of weekly recurrent CBT using X-Ray Tutor (XRT). A control group of screeners was tested as well but this group did not receive training with XRT. Large increases in detection performance were found for the training group, which did also generalize to new threat objects that were not shown during training. The results of this study indicate that recurrent CBT can be a powerful tool to increase the x-ray image interpretation competency of screeners.

4.1.2 Introduction

In recent years, x-ray screening of passenger bags has become an essential component of airport security. Large investments were made into state-of-the art x-ray screening equipment. However, well trained human screeners are needed to operate the equipment appropriately in order to detect threat objects in passenger luggage within few seconds of inspection time. Object shapes that are not similar to ones stored in visual memory are difficult to recognize (e.g., Graf, Schwaninger, Wallraven, & Bülthoff, 2002; Schwaninger, 2004, 2005). Thus, a prerequisite for good threat detection performance is knowledge about which objects are prohibited and what they look like in x-ray images. Schwaninger, Hardmeier, and Hofer (2005) have shown that x-ray screener performance depends on knowledge-based and image-based factors. Image-based factors refer to image difficulty resulting from viewpoint variation of threat objects, superposition of threat objects by other objects in a bag, and bag complexity depending on the number and type of objects in the bag. The ability to cope with image-based factors is related to individual visual-cognitive abilities rather than a mere result of training. In contrast, knowledge-based factors refer to knowing which items are prohibited and what they look like in x-ray images of passenger bags. Because objects look quite different in x-ray images than in reality and because many threat objects are not known from everyday experience, computer-based and on the job training are important determinants of x-ray detection performance. Schwaninger et al. (2005) compared detection performance of novices with the one of trained aviation security screeners. A rather poor recognition of unfamiliar object shapes (e.g. self-defense gas spray, electric shock device etc.) in x-ray images was found for novices. For trained aviation security personnel, a much higher recognition performance was shown. Schwaninger and Hofer (2004) showed that adaptive computer-based training (CBT) can be very effective to increase the detection of improvised explosive devices (IEDs) in x-ray images of passenger bags. McCarley, Kramer, Wickens, Vidoni, and Boot

(2004) reported a better performance after training for the detection of knives in x-ray images.

The purpose of this study was to investigate to what extent the previous findings can be expanded to other threat categories (e.g., guns and other prohibited items) and to examine transfer effects. The training group conducted weekly recurrent CBT using X-Ray Tutor (Schwaninger, 2004). The control group did not receive this type of training and conducted recurrent classroom training including another CBT system. Both groups of screeners were tested before and after 6 months using the X-Ray Competency Assessment Test (X-Ray CAT, Koller & Schwaninger, 2006). This test shows different kinds of prohibited items in x-ray images of passenger bags. Half of the threat objects in the X-Ray CAT were not presented during the training sessions. This enabled measuring whether a transfer of the gained knowledge about trained objects to untrained but similar looking objects occurs.

4.1.3 Method

4.1.3.1Participants

A total of 209 airport security screeners of a European airport participated in this study and conducted the X-Ray CAT 1.0.0 two times with an interval of six months. The training group consisted of 97 screeners who conducted weekly recurrent CBT of about 20 minutes using X-Ray Tutor (XRT) CBS 2.0 Standard Edition during the 6 months interval between the two test measurements. The control group consisted of 112 screeners and they did not conduct weekly recurrent CBT with XRT.

4.1.3.2 Materials

The X-Ray CAT consists of 128 x-ray images of passenger bags. Each of the bags is used twice, once containing a prohibited item (threat image) and once without any threat object (Figure 8 displays an example of the stimuli). The threat items belong to four categories of prohibited items as defined in Doc 30 of the European Civil Aviation Conference (ECAC): guns, improvised explosive devices (IEDs), knives and other prohibited items (e.g., gas, chemicals, grenades etc.). The threat objects have been selected and prepared by experts of Zurich State Police, Airport division to be representative and realistic.

Figure 8: Example of an x-ray image of a passenger bag. The image on the right contains the prohibited item depicted separately on the bottom right.

For each threat category 16 exemplars are used (8 pairs). Each pair consists of two prohibited items that are similar in shape (see Figure 9). These were distributed randomly into two sets, set A and set B.

Figure 9: Example of two x-ray images of similar looking threat objects used in the test, one belonging to set A and B, respectively.

Every item is depicted from two different viewpoints. The easy viewpoint shows the object from a canonical perspective (Palmer, Rosch, & Chase, 1981) as judged by two security experts who captured the stimuli. The difficult viewpoint shows the threat item with an 85 degree horizontal rotation or an 85 degree vertical rotation relative to the canonical view. In each threat category half of the prohibited items of the difficult viewpoint are rotated vertically, the other half horizontally. Set A and B are equalized concerning the rotations of the prohibited objects. The effects of viewpoint are not analyzed in this study and will be reported elsewhere.

Every threat item is combined with a bag in a manner that the degree of superposition by other objects is similar for both viewpoints. This was achieved using a function that calculates the difference between the pixel intensity values of the bag image with the threat object minus the bag image without the threat object using the following formula (see also section 3.1.11.3):

$$SP = \frac{\sqrt{I_{SN}(x,y) - I_N(x,y)}}{ObjectSize}$$

SP = Superposition; I_{SN} = Grayscale intensity of the SN (Signal plus Noise) image (contains a prohibited item); I_N = Grayscale intensity of the N (Noise) image (contains no prohibited item); Object Size: Number of pixels of the prohibited item where R, G and B are < 253

Using this equation (division by object size), the superposition value is independent of the size of the prohibited item. This value can be kept relatively constant for the two views of a threat object, independent of the degree of clutter in a bag, when combining the bag image and the prohibited item. The bag images were visually inspected by aviation security experts to ensure they do not contain any other prohibited items. Harmless bags were assigned to the different categories and viewpoints of the threat objects in a way that their difficulty was balanced across all categories[1]. The false alarm rate (the rate at which screeners wrongly judged a harmless bag as containing a threat item) for each bag image served as measure of difficulty based on a pilot study with 192 screeners.

The X-Ray CAT is integrated in the XRT training system and takes about 20-30 minutes to complete. Each image is shown for a maximum of 10 seconds on the screen. Screeners have to judge whether the bag is OK (contains no prohibited item) or NOT OK (contains a prohibited item). Additionally, screeners have to indicate the perceived difficulty of each image on a 100 point scale (difficulty rating). The difficulty ratings were not analyzed in study and will be reported elsewhere. The visible appearance of the test is the same as in training except there is no feedback and screeners do not have to click on the image to identify the threat object (see Figure 10). Feedback is provided only during training and informs the screener whether the image has been judged correctly or not. If the bag contains a threat item, it is highlighted by flickering after the screener responded with OK or NOT OK and the screener has the possibility to display information about the threat item (see Figure 10). As mentioned previously, during training, screeners have to click on the image and mark the object they perceive to be a threat item. This is not required during test mode.

[1] The eight categories of test images (four threat categories in two viewpoints each) are similar in terms of the difficulty of the harmless bags. This means, a difference of detection performance between categories or viewpoints cannot be due to differences in the difficulty of the bag images.

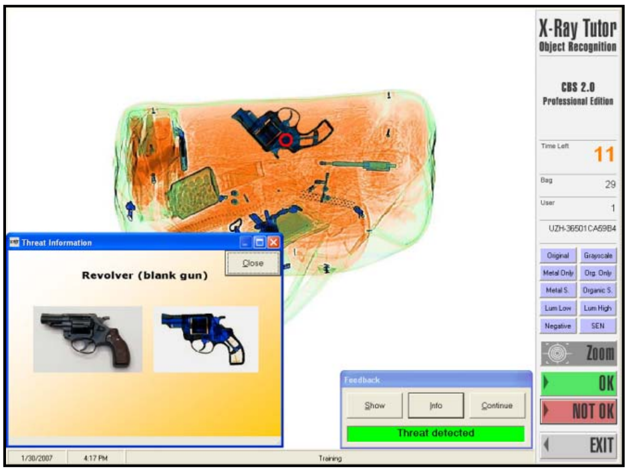

Figure 10: Screenshot of the XRT training system during training. At the bottom right a feedback is provided. If a bag contains a threat item, an information window can be displayed (see bottom left of the screen).

4.1.4 Procedure

Screeners were randomly distributed into two groups. Both groups conducted the X-Ray CAT 1.0.0 without having trained with XRT before (baseline measurement). After test completion, only one group received recurrent adaptive CBT using XRT (training group). On average, each screener of the training group conducted 20.26 min recurrent training per week (SD = 3.65 min). After six months, both groups conducted the X-ray CAT again. This approach allows the comparison of the two test measurements and the performance of the two groups prior to and after training with XRT.

In order to measure a transfer effect, only the images of the prohibited items of test set A were included in training. They are part of the XRT CBS 2.0 SE training library, which contains 100 threat items belonging to the four threat categories (guns, IEDs, knives, other). Most of them are depicted from six different viewpoints. No bag image of the test appeared during training with XRT. During training, images containing a threat object are created at the point of use, that is, test threat items (set A) and other threat items are digitally inserted into randomly selected bag images at random positions. For details on XRT see Schwaninger (2004).

4.1.5 Results

Detection performance was calculated using the signal detection measure d’ (Green & Swets, 1966), which takes into account the hit rate (correctly judged threat images as being NOT OK) and the false alarm rate (wrongly judged harmless bags as being NOT OK). Figure 11 shows the detection performance of the first and second measurement for both screener groups. Performance values are not reported due to security reasons. However, effect sizes are reported for all relevant analyses and interpreted based on Cohen (1988), see Table 6.

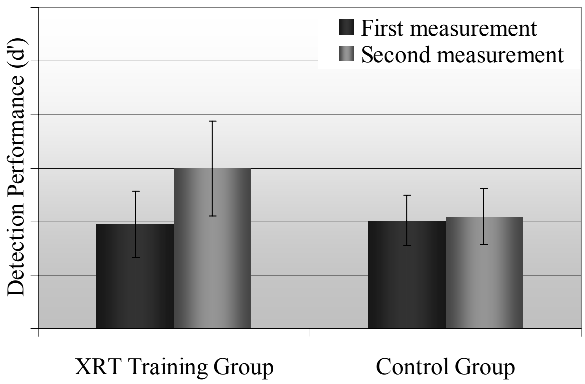

Figure 11: Detection performance with standard deviations for the XRT training group vs. the control group comparing first and second measurement.

Table 6: Classification of effect sizes according to Cohen (1988)

Effect size	d	η^2
small	0.20	0.01
medium	0.50	0.06
large	0.80	0.14

An analysis of variance (ANOVA) for repeated measures using d’ scores with the within-participant factor measurement (first vs. second) and the between-participant factor group (trained vs. control) revealed a large main effect of measurement (first vs. second), $\eta^2 = .40$, $F(1, 207) = 138.66$, $p < .001$, a medium main effect of group (trained vs. control), $\eta^2 = .13$, $F(1, 207)$

= 31.22, p < .001, and a large interaction of measurement and group η^2 = .34, F(1, 207) = 105.55, p < .001.

Separate pairwise t-tests of detection performance d' revealed no significant difference at the baseline measurement between the two groups (p = .353) and no significant difference for the control group in both measurements (p = .108). However, there was a significant difference for the XRT training group between the first and the second measurement (p < .001) with a large effect size of d = 1.39. There was also a significant difference between the two groups at the second measurement, p < .001, with a large effect size of d = 1.27.

Figure 12 shows the detection performance for each threat category separately for both groups at the first and the second measurement. A repeated-measures ANOVA with the within-participant factors measurement (first vs. second) and threat category (guns, IEDs, knives and other), and the between-participant factor group (XRT training vs. control) revealed the main effects and interactions given in Table 2a.

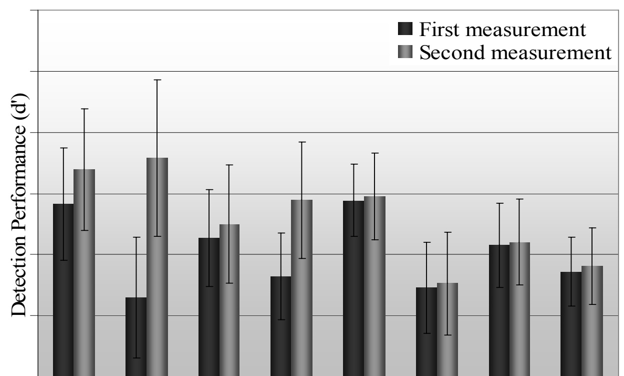

Figure 12: Detection performance with standard deviations for the XRT training group vs. the control group comparing first and second measurement for each threat category separately.

Separate pairwise t-tests were conducted to compare detection performance at the first and the second measurement for both groups and each threat category separately. The XRT training group showed a significant increase of the detection performance at the second measurement for each threat category (guns, IEDs and other threat objects, all p < 001, all d > 0.60, knives, p < .05, d = 0.26). Detection performance of the control group did not differ significantly between the two measurements (guns: p = .358, IEDs: p = .296, knives: p = .467, and other threat objects: p = .168).

The results of the analysis considering the two sets of the test, set A and set B, are shown in Figures 13 and 14. The results of the repeated measures ANOVA with the within-participant factors measurement (first vs. second) and test set (A vs. B) and the between-participant factor group (XRT training group vs. control group) can be seen in Table 7b. Pairwise t-tests showed a significant increase in detection performance at the second measurement for both sets for the XRT training group (set A and B: p < .001, d > 1.25) but not for the control group.

Table 7: Results of the ANOVAs

	Factor	df	F	η^2	p
a)	Measurement (M)	1, 207	140.23	0.40	<.001
	Threat Category (T)	3, 621	222.7	0.52	<.001
	Group (G)	1, 207	37.57	0.15	<.001
	M x G	1, 207	108.16	0.34	<.001
	T x G	3, 621	29.36	0.12	<.001
	M x T	3, 621	76.5	0.27	<.001
	M x T x G	3, 621	74.78	0.27	<.001
b)	Measurement (M)	1, 207	138.39	0.40	<.001
	Group (G)	1, 207	32.64	0.14	<.001
	Test Set (S)	--	--	--	n.s.
	M x G	1, 207	104.08	0.34	<.001
	M x S	1, 207	8.72	0.04	<.01
	S x G	1, 207	17.31	0.08	<.001
	M x S x G	1, 207	7.92	0.04	<.01
c)	Measurement (M)	1, 207	146.15	0.41	<.001
	Threat Category (T)	3, 621	219.54	0.52	<.001
	Group (G)	1, 207	42.53	0.17	<.001
	M x G	1, 207	108.68	0.34	<.001
	T x G	3, 621	30.29	0.13	<.001
	M x T	3, 621	78.18	0.27	<.001
	M x S	1, 207	10.17	0.05	<.01
	T x S	3, 621	58.12	0.22	<.001
	M x T x G	3, 621	75.51	0.27	<.001
	M x S x G	1, 207	6.67	0.02	<.05

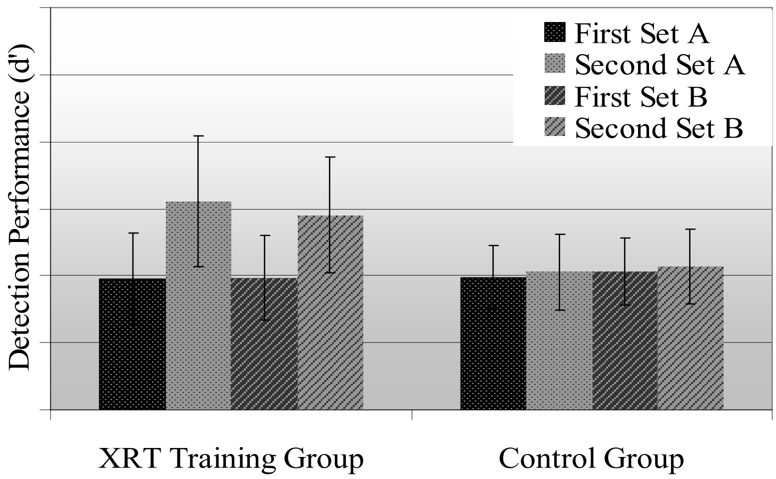

Figure 13: Detection performance with standard deviations for the XRT training group vs. the control group comparing first and second measurement for set A and set B separately.

An ANOVA for repeated measures with the within-participant factor set showed a very small significant main effect of set $\eta^2 = .02$, $F(1, 208) = 3.94$, $p < .05$ at the first measurement. Pairwise *t*-tests comparing both sets within one group at the first measurement revealed a significant difference of the two sets only for the control group with only a small effect size ($p < .01$, $d = 0.17$) but not for the XRT training group ($p = .676$).

An extended ANOVA with the additional within-participant factor threat category revealed the main effects and interactions as specified in Table 7c.

Pairwise *t*-tests confirmed a significant ($p <$ 001, all $d > 0.46$) increase in detection performance for the XRT training group for all threat categories per set except for knives (set A: $p < .05$, $d = 0.27$, set B; $p = .127$, $d = 0.19$). The control group showed no significant change in detection performance at the second measurement for neither threat category per set (set A: guns $p = .147$, IEDs $p = .202$, knives $p = .801$, other threat objects $p = .245$; set B: guns $p = .974$, IEDs $p = .597$, knives $p = .235$, other threat objects $p = .123$).

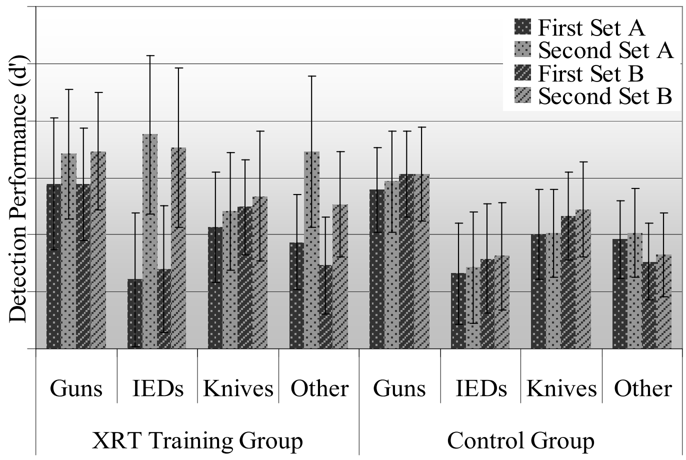

Figure 14: Detection performance with standard deviations for the XRT training group vs. the control group comparing first and second measurement for set A and set B and each threat category separately.

4.1.6 Discussion

The results of this study confirmed earlier findings on x-ray detection performance of airport security screeners showing that adaptive CBT with X-Ray Tutor (XRT) results in substantial increases of detection performance (e.g., Hardmeier, Hofer, & Schwaninger, 2006; Schwaninger & Hofer, 2004; Schwaninger et al., 2005). In this study, the training group showed remarkable increases in detection performance for all types of threat objects (guns, knives, IEDs, and other prohibited items). For the control group, which did not conduct weekly recurrent CBT with XRT, no significant change in detection performance was observed. It should be noted that according to the security organization and their appropriate authority, the control group did recurrent training as mandated by national regulation during the whole duration of the study. This training was comparable in terms of the required training hours and included x-ray image interpretation training using another commercially available CBT. Thus, the improved performance in the training group reflects specific effects of training with XRT and they cannot be explained by a "Hawthorne Effect". The largest training effect was found for IEDs. It should be noted that as all other stimuli, the IEDs were developed by police experts of Zurich State Police, Airport Division. Especially the IEDs were quite sophisticated threat objects using components that are often not known to screeners without enhanced training in IED detection. Thus it is not surprising that before training, d' scores for IEDs were substantially smaller than for guns. However, after training, IED detection of the training group was very good and even slightly better than gun detection. This shows that the detection of IEDs is not difficult per se, but rather depending on the training of screeners.

Besides measuring training effects, the main aim of this study was to examine whether gained knowledge about trained threat objects can be transferred to similar looking objects. Since the X-Ray CAT is composed of two comparable sets (set A and set B) this can easily be tested by including the threat objects of one set (in this case set A) into the XRT system. A large transfer effect would mean a similarly higher detection performance after training for both sets. This was confirmed, as Figures 13 and 14 illustrate. The significant increase of the detection performance for the XRT training group was found for the trained test set A as well as for the untrained test set B. This implies a large transfer of the acquired knowledge about the visual appearance of trained objects (set A) to untrained but similar looking objects (set B). The comparison of the two sets A and B at the baseline measurement over all screeners showed a slightly significant difference ($p < .05$) indicating that the two sets are not exactly equal in terms of image difficulty. But this possible objection to the transfer effect can be disapproved with two arguments: first, the effect size is only small according to the conventions by Cohen (1988, see also Table 1), and second, only the control group showed a significant difference ($p < .01$) but not the XRT training group ($p = .676$). Therefore, the transfer effect in the results of the XRT training group can be attributed to the training of set A only.

Transfer effects were revealed for all threat categories, i.e. for guns, IEDs and other threat objects. For knives, a significant training effect appeared only in the trained set A ($p < .05$) but not in the untrained set B ($p = .127$). Thus, there was no transfer effect for knives from set A to set B. Either the knives of the two sets were not similar enough in shape to allow a transfer effect, or the small training effect for knives is due to their shape. On one hand, knives show less diagnostic features which play an important role in object recognition compared to objects from other categories. On the other hand, the visual similarity of knives to harmless everyday objects (e.g., pen) is substantial. These factors could impede detectability and trainability and ultimately might have resulted in small transfer effects.

Contrary to our results, Smith, Redford, Gent, and Washburn (2005) found a large decrease in screeners' detection performance when specific trained objects were replaced with new images belonging to the same categories (see also Smith, Redford, Washburn, and Taglialatela, 2005). According to these authors, improvement in screening performance is attributable only to specific-token familiarity that developed for the original images and not to a category generalization. They state constraints on categorization and the use of category-general information when humans face visual complexity and have to identify targets within it. Our results can be interpreted in support of generalization of visual learning in x-ray image interpretation. However, it might be possible that the objects of the untrained set in our study are so similar to the trained objects that a specific-token familiarity led to the detection performance increase and not a true generalization effect.

The lacking transfer effect in knives would along these lines mean that the objects in set A and set B are not similar enough in shape to generate a specific-token familiarity. Therefore only the learnt objects could generate a training effect but not the unlearnt ones. For Schwaninger and Hofer's (2004) findings of a large increase in detection performance of IEDs after recurrent CBT with other members of the category than those included in the test, it would mean, that those objects were very similar in order to create a specific-token familiarity and therefore a training effect.

For our future studies, it could also be interesting to increase the interval between the end of training and the testing of training transfer, as corresponding literature usually tests transfer of training after a considerable period of time in order to measure the stability of the transfer (e.g., Saks & Belcourt, 2006). However, most research is about organizational training and therefore training transfer is related to learning working skills and the generalization to the job context (Baldwin & Ford, 1988). In contrast our transfer refers to the transfer of visual knowledge about objects to other objects.

In any case, our findings show that the knowledge about the visual appearance of forbidden objects, which airport security screeners acquire during recurrent CBT, can be transferred to similar looking, but not previously seen objects. Thus, adaptive CBT can be a powerful tool to increase screeners' x-ray image interpretation competency.

4.1.7 Acknowledgments

This research was financially supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403). Many

thanks to Zurich State Police, Airport Division, for their help in creating the stimuli and the good collaboration for conducting parts of the study.

4.1.8 References

Baldwin, T., & Ford, J.K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41,* 63-105.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences.* New York: Erlbaum, Hillsdale.

Graf, M., Schwaninger, A., Wallraven, C., & Bülthoff, H.H. (2002). Psychophysical results from experiments on recognition & categorisation. *Information Society Technologies (IST) programme, Cognitive Vision Systems – CogVis (IST-2000-29375)*.

Green, D.M., & Swets, J.A. (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley.

Hardmeier, D., Hofer, F., & Schwaninger, A. (2006). The role of recurrent CBT for increasing aviation security screeners' visual knowledge and abilities needed in x-ray screening. *The 4th International Aviation Security Technology Symposium, Washington, D.C., USA, November 27 – December 1, 2006.*

Koller, S., & Schwaninger, A. (2006). Assessing X-ray image interpretation competency of airport security screeners. *Proceedings of the 2nd International Conference on Research in Air Transportation, ICRAT 2006 Belgrade, Serbia and Montenegro, June 24-28, 2006,* 399-402.

McCarley, J.S., Kramer, A.F., Wickens, C.D., Vidoni, E.D., & Boot, W.R. (2004). Visual skills in airport screening. *Psychological Science, 15 (5),* 302-306.

Palmer, S.E., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In I. Long and A. Baddeley (Eds.), *Attention and Performance IX*. Hillsdale, N.J.: Erlbaum.

Saks, A.M, & Belcourt, M. (2006). An Investigation of Training Activities and Transfer of

Training in Organizations. *Human Resource Management, 45 (4),* 629-648.

Schwaninger, A. (2004). Computer based training: a powerful tool to the enhancement of human factors. *Aviation Security International, FEB/2004,* 31-36.

Schwaninger, A. (2005). Object recognition and signal detection. In B. Kersten (Ed.), *Praxisfelder der Wahrnehmungspsychologie*. Bern: Huber.

Schwaninger, A., Hardmeier D., & Hofer F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. *IEEE Aerospace and Electronic Systems, 20(6)*, 29-35.

Schwaninger, A., & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in x-ray screening. In K. Morgan and M.J. Spector, *The Internet Society 2004, Advances in Learning, Commerce and Security.* Wessex: WIT Press.

Smith, J.D., Redford, J.S., Gent, L.C., & Washburn, D.A. (2005). Visual search and the collapse of categorization. *Journal of Experimental Psychology: General, 134 (4),* 443-460.

Smith, J.D., Redford, J.S., Washburn, D.A., & Taglialatela, L.A. (2005). Specific-token effects in screening tasks: possible implications for aviation security. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31 (6),* 1171-1185.

4.2 Computer-based training increases efficiency in x-ray image interpretation by aviation security screeners

4.2.1 Abstract

X-Ray screening of passenger bags is an essential component of airport security. Large investments into technology have been made in recent years. However, the most expensive equipment is of limited value, if the humans who operate it are not selected and trained appropriately. Scientific studies have shown that human performance in x-ray image interpretation depends critically on individual abilities and visual knowledge acquired through experience on the job and training. The aim of this study was to investigate the effect of adaptive computer-based training for increasing the detection of guns, knives, improvised explosive devices (IEDs), and other prohibited items. 97 airport security screeners of a European airport participated in this study. At the beginning of the project, all airport security screeners conducted the X-Ray Competency Assessment Test (X-Ray CAT). Thereupon they received adaptive computer-based training (CBT) for about 4 months. Then they conducted the X-Ray CAT the second time in the middle of the project. This was followed by about 4 months of CBT and a third test with X-Ray CAT at the end of the project. The goal was that each screener conducts at least one 20 minute training session per week. Substantial increases of detection performance were found as a result of training, which depended on the threat category (guns, IEDs, knives and other prohibited items). The largest training effects were found for IEDs. Additional analyses showed that training not only leads to an increase of detection performance but also results in faster response times when an x-ray image contains a threat

object. Thus, recurrent CBT can be a powerful tool to increase efficiency in x-ray image interpretation by airport security screeners.

4.2.2 Introduction

In response to the increasing threat of terrorist attacks, large investments were made into x-ray screening machines of the newest generation in order to inspect passenger baggage at airport security checkpoints. The last decision however is always made by a human operator (screener). The most expensive equipment is of little value if the screeners who operate it are not selected and trained appropriately. They have to be able to detect threat objects in passenger luggage within few seconds of inspection time.

Object shapes that are not similar to ones stored in visual memory are difficult to recognize (Graf, Schwaninger, Wallraven & Bülthoff, 2002; Schwaninger, 2004a; Schwaninger 2004b). Detection of forbidden objects in x-ray images of passenger baggage depends on knowledge-based as well as on image based factors (Schwaninger, Hardmeier & Hofer, 2005). An airport security screener has to know which objects are prohibited and what they look like in an x-ray image. Some objects look quite different in x-ray images than in reality, for example an electric shock device. Other threat objects, like improvised explosive devices (IEDs), are rarely seen, in every day life as well as at the security checkpoint. Thus, it is not surprising that computer-based training (CBT) is very important to achieve and maintain a high detection performance, which is of special importance for detecting IEDs (Hardmeier, Hofer & Schwaninger, 2006; Schwaninger & Hofer, 2004). Furthermore, threat objects can be depicted in an unfamiliar rotation in the baggage which can have a big impact on the detection performance. Based on research findings from object recognition by Schwaninger (2004) a large and representative image library of prohibited items depicted from different viewpoints is necessary to provide a good basis for training x-ray image interpretation competency. In addition to knowledge based factors, also image based factors play an important role. These can rather be attributed to the visual abilities of a person, that is, to the abilities to cope with image difficulty resulting from rotation of a threat object, superposition by other objects in the bag, and bag complexity (Schwaninger, Hardmeier &. Hofer, 2005; Hardmeier, Hofer & Schwaninger, 2005; Hardmeier, Hofer & Schwaninger, 2006; Hofer, Hardmeier & Schwaninger, 2006; Schwaninger, 2003).

A comparison of the detection performance of novice screeners with the one of trained aviation security screeners in an earlier study revealed a rather poor recognition of unfamiliar object shapes (e.g. self-defence gas spray, electric shock device etc.) in x-ray images for novices, whereas for trained aviation security personnel a much higher recognition performance was shown (Schwaninger, Hardmeier & Hofer, 2005). Schwaninger & Hofer (2004) showed that adaptive CBT can be very effective to increase the detection of improvised explosive devices (IEDs) in x-ray images of passenger bags. McCarley, Kramer, Wickens, Vidoni & Boot (2004) reported a better performance after training for the detection of knives in x-ray images.

4.2.3 Method and Procedure

The aim of this research project was to investigate to what extent recurrent adaptive CBT using X-Ray Tutor increases x-ray image interpretation competency. 97 airport security screeners of a European airport participated in this study. At the beginning of the project all airport security screeners conducted the X-Ray Competency Assessment Test (X-Ray CAT, Koller & Schwaninger, 2006). Thereupon they received training for about 4 months, then conducted the X-Ray CAT a second time in the middle of the project. This was followed by about 4 months of training and a third test with X-Ray CAT at the end of the project. The goal was that each screener conducts at least one 20 minute training session per week.

4.2.3.1X-Ray Competency Assessment Test (X-Ray CAT)

The X-Ray CAT is a standardized, reliable and valid instrument to measure x-ray image interpretation competency as defined by the principles and requirements specified in Schwaninger, Bridges, Drury, Durinckx, Durrant, Hodge, Hofer, Jongejan, Maguire, McClumpha, Neiderman, Steinmann & Wüest (2005). It contains 256 x-ray images of passenger baggage (see Figure 15).

Figure 15: Example of an x-ray image of a passenger bag. The image on the right contains the prohibited item depicted separately on the bottom right.

Half of the bags contain a threat item, the other 128 bags are harmless. The threat items belong to the four categories guns, knives, IEDs and other prohibited items as defined in ECAC DOC 30. Each category is represented by 16 threat objects (8 visually similar pairs) and each object is depicted in the baggage in an easy and a difficult view. Of the visually similar pairs, only one item is used in training with X-Ray Tutor, while the other item is not used during training. A recent study showed that the performance improvements as a result of training with X-Ray Tutor generalize to visually similar objects not shown during training (Koller, Hardmeier, Michel, & Schwaninger, 2007). While the easy view corresponds to the most usual (canonical) view, in the difficult view the threat object is rotated 85° either around the horizontal or the vertical axis. At test, airport security screeners have to decide for each bag whether it is OK (bag without threat item) or NOT OK (bag containing a threat item). Each image is depicted for a maximum of 15 seconds. Depending on how many images an airport security screener can visually inspect during 20 minutes, the test lasts about 2-3 sessions of 20 minutes. For more detailed information about the X-Ray CAT (see Koller & Schwaninger, 2006; Koller, Hardmeier, Michel & Schwaninger, 2007).

4.2.3.2 X-Ray Tutor

X-Ray Tutor is a scientifically based training program. It is based on findings about how the human brain processes visual information in order to recognise objects in different views, when superimposed by other objects, and depending on bag complexity (Schwaninger, 2003). The training is individually adaptive, that is, it automatically adapts to the performance of individual airport security screeners. X-Ray Tutor automatically combines images of fictional threat items with x-ray images of passenger bags. This is performed by an individually adaptive algorithm, which takes into account the rotation of threat objects, the superposition by other objects in the bag, and bag complexity resulting from clutter and transparency of the objects in the baggage. X-Ray Tutor 2.0 contains a large image library of threat objects that are depicted in different standardized views. Most of the objects can be depicted from up to 72 different viewpoints, which allows training screeners to detect threat objects independent of rotation. This image library was built in close collaboration with experts of Zurich State Police, Airport Division, and it is being extended continuously.

During training with X-Ray Tutor, x-ray images of bags are depicted on the screen for 15 seconds (standard setting). Screeners have to decide whether the bag is OK (i.e. it contains no threat object) or whether it is NOT OK (i.e. it contains a threat object). After each response, a feedback is provided informing the screener whether his/her response was correct. If the bag contained a threat object the user can view detailed information and a real image of the threat object. For further information on X-Ray Tutor see Schwaninger, 2004.

4.2.4 Results and Discussion

In this study, the performance of screeners to detect threat objects in the X-Ray CAT as well as the time needed for detecting the threat objects (i.e. reaction time) has been analyzed. An effect of training could imply an increase in detection performance and/or a decrease in reaction time. Furthermore, the effect of object viewpoint has been analyzed, i.e. a possible difference in detection performance of threat objects depending on the rotation with which they are depicted in the image.

4.2.4.1 Detection Performance

Detection performance in the X-Ray CAT was analysed using d', a widely used measure of sensitivity based on signal detection theory (Green & Swets, 1966; Hofer & Schwaninger, 2004; MacMillan & Creelman, 1991). The d' measure takes into account the hit rate as well as the false alarm rate. It can be calculated by the following formula: $d' = z(H) - z(FA)$, whereas H is the hit rate, FA the false alarm rate and z refers to the z-transform.

The hit rate indicates how often a person correctly judges a bag as being NOT OK proportionately to all bags containing a threat object. The false alarm rate indicates how often a person wrongly judges a bag as being NOT OK proportionately to all bags containing no threat object. In this study, actual performance values are not reported due to security reasons. However, effect sizes are reported for all relevant analyses and interpreted based on Cohen (1988).

4.2.4.1.1 Effect of Training

Figure 16 shows the detection performance and the standard deviation[2] for the easy view of the threat objects in each category for all three test measurements. Guns were detected best in all three tests and objects of the categories IEDs and "Other" were detected worst. Substantial increases of detection performance were found, which depended on the threat category. The largest training effects were found for IEDs. Good performance was achieved for IEDs after the two training blocks of 4 months each. The aim was that all screeners conduct at least one 20 minute training session per week, which was achieved on average. Note

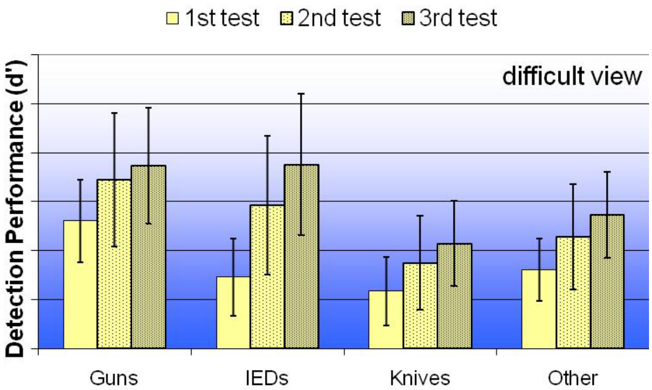

Figure 16: Detection performance d' and standard deviations for difficult views broken up by threat category and test date. *Note*: For security reasons

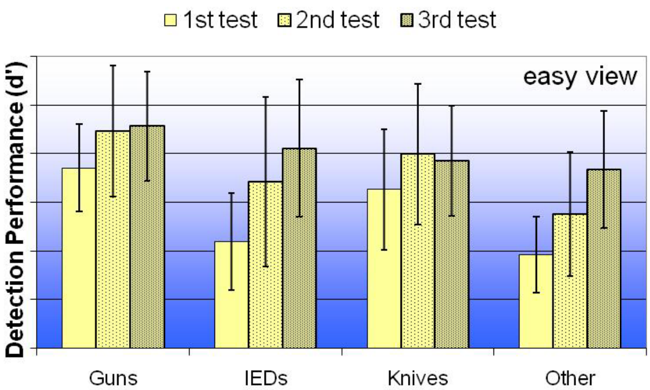

Figure 17: Detection performance d' and standard deviations for easy views broken up by threat category and test date. *Note*: For security reasons d' scores are not indicated in the figure.

[2] The standard deviation represents the range of dispersion around the mean of the data and indicates the range of individual differences between the tested airport security screeners.

that there are large differences between screeners as can be seen by the large standard deviations (thin lines in Figure 16). Some screeners achieved very good performance for all types of threat objects after the two training phases. This is mainly due to differences in the amount of training. While some screeners did only a few trainings over several months, other screeners did several training sessions per week and achieved very high performance increases.

An analysis of variance (ANOVA) for repeated measures using d' scores for easy view with the within-participant factors test date (first, second, third) and category (guns, knives, IEDs, other) revealed large main effects of test date $\eta^2 = .36$, $F(2, 192) = 53.82$, $p < .001$, and category $\eta^2 = .57$, $F(3, 288) = 128.54$, $p < .001$, and a large two-way interaction of test date and category $\eta^2 = .15$, $F(6, 576) = 16.80$, $p < .001$. These results confirm that CBT with X-Ray Tutor result in large performance increases of x-ray screeners, especially regarding the detection of IEDs and other threat items.

4.2.4.1.2 Effect of Object Viewpoint

Figure 18 shows the results for detection performance d’ and standard

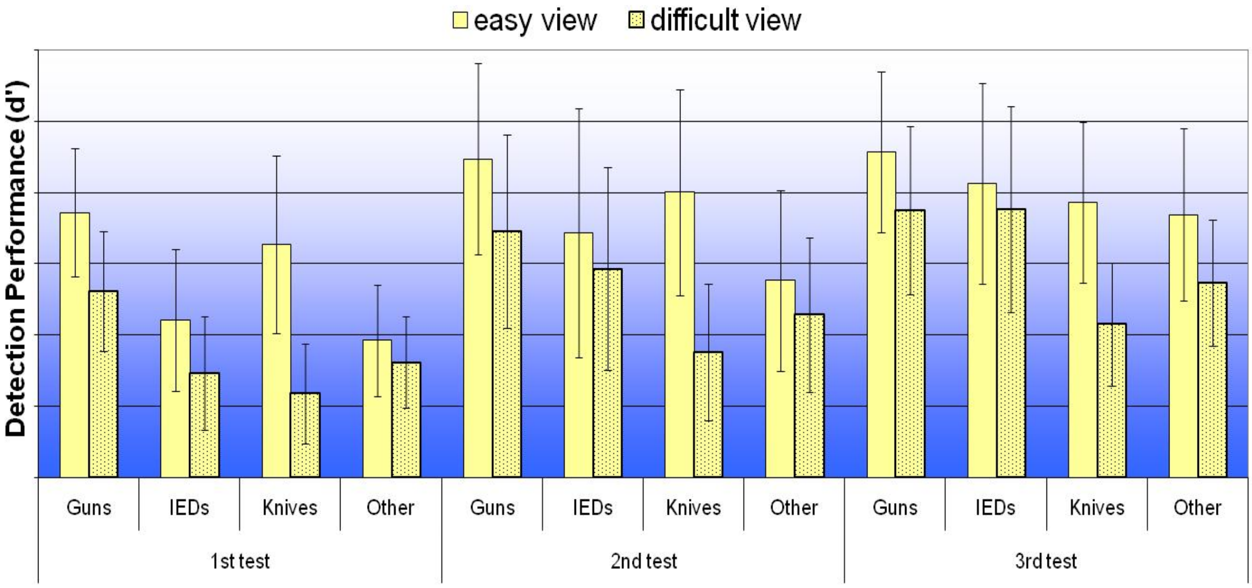

Figure 18: Detection performance d’ and standard deviations broken up by threat category, view (easy vs. difficult), and test date. *Note*: For security reasons d’ scores are not indicated in the figure.

deviations when threat objects were depicted from a difficult viewpoint. Figure 18 depicts the comparison between the detection performance for objects in easy view and difficult view. It shows that detection of threat objects is much easier in frontal (or canonical view) than when depicted from a difficult viewpoint. Large training effects have been found for difficult views (cf. knives) as well as for objects that are rarely seen in every day life (IEDs and other threat items). The reason for this large performance increase is the fact that X-Ray Tutor contains a large threat image library in which objects are depicted in many different viewpoints (Schwaninger, 2004). X-Ray Tutor trains each screener individually to become able to detect all types of threat objects even if they are shown from a difficult viewpoint. This is the reason why such large training effects were found for objects shown in difficult view.

An analysis of variance (ANOVA) for repeated measures using d' scores for difficult view with the within-participant factors test date (first, second, third) and category (guns, knives, IEDs, other) revealed a large main effect of test date $\eta^2 = .46$, $F(2, 192) = 82.22$, $p < .001$, a large main effect of category $\eta^2 = 68$, $F(3, 288) = 200.32$, $p < .001$, and a large two-way interaction of test date and category $\eta^2 = .15$, $F(6, 576) = 16.94$, $p < .001$. These results provide further evidence for the effectiveness of recurrent CBT with X-Ray Tutor, which results in large performance increases for detecting threat items in x-ray images.

An additional analysis of variance (ANOVA) for repeated measures using d' scores for both views (easy and difficult) with the within-participant factors test date (first, second, third), view (easy vs. difficult), and category (guns, knives, IEDs, other) was conducted to examine the effect of viewpoint on the detection performance of screeners. There was a large main effect of test date $\eta^2 = .43$, $F(2, 192) = 73.23$, $p < .001$, a large main effect of view (easy vs. difficult) $\eta^2 = .89$, $F(1, 96) = 804.15$, $p < .001$, and a large main effect of category $\eta^2 = .66$, $F(3, 288) = 182.39$, $p < .001$. The following interactions were significant as well with large effects for the two-way interactions between test date and category $\eta^2 = .19$, $F(6, 576) = 22.34$, $p < .001$, and between viewpoint and category $\eta^2 = .56$, $F(3, 288) = 120.00$, $p < .001$. There was also a medium effect for the three-way interaction of test date, view and category, $\eta^2 = .08$, $F(6, 576) = 7.87$, $p < .001$.

These results show that recurrent CBT with X-Ray Tutor is very effective to train screeners to detect threat objects even if they are depicted from an unusual viewpoint.

4.2.4.1.3 Reaction Times

For each response, reaction time (RT) was measured, i.e. the time between x-ray image onset and the time a response was provided by the screener (OK or NOT OK button). Figure 19 shows the RTs of all hits (correctly judged images as NOT OK) of all three test dates broken up by threat category. RTs decreased as a result of training, especially from the first to the second test. However, there were also large differences between individual airport security screeners (cf. large standard deviations), possibly due to differences in amount of training.

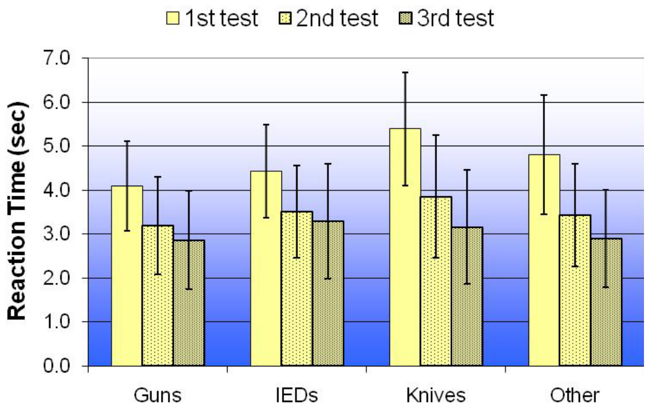

Figure 19: Reaction times and standard deviations of hits (correctly answered as NOT OK) broken up by threat category and test date.

An analysis of variance (ANOVA) for repeated measures using RTs with the within-participant factors test date (first, second, third) and category (guns, knives, IEDs, other) revealed a large main effect of test date $\eta^2 = .47$, $F(2, 192) = 85.01$, $p < .001$, a large main effect of category $\eta^2 = .44$, $F(3, 288) = 76.42$, $p < .001$, and a large two-way interaction of test date and category $\eta^2 = .16$, $F(6, 576) = 18.79$, $p < .001$.

4.2.5 Summary and Conclusion

This study has shown substantial increases of airport security screeners' x-ray image interpretation competency as a result of recurrent adaptive CBT using X-Ray Tutor. The largest increase was found for the detection of IEDs. After two training phases of 4 months each, detection was almost as good as detection of guns. There were also large effects of viewpoint. Objects shown from a difficult rotation are more difficult to recognize (especially knives) than when depicted from a non-rotated canonical view. However, the effect of viewpoint can be compensated by training. At the third test, the difficult views were recognized much better than before training started (first test). There were also large differences between screeners. While some screeners did only a few trainings per month, others did several training sessions per week and achieved very large performance increases. A large effect of training was also found in the reaction times. Screeners could reduce the time needed to detect a threat object significantly. More detailed analyses showed that the increase in the detection performance was mainly due to an increase in the hit rate (as opposed to a decrease in the false alarm rate), which means, that a speed-accuracy trade off can be ruled out here.

It is not surprising that before training, d' scores for IEDs and also for other threat categories were substantially smaller than for guns. The IEDs used in this study are quite sophisticated threat objects using components that are often not known to screeners without enhanced training in IED detection. For other threat items probably this knowledge based factor comes into play as well. Screeners first have to learn which objects are prohibited and what they look like in x-ray images. If an effective CBT is used for recurrent training, a large increase in detection performance can be achieved (see also Schwaninger, 2004). This shows that the detection of IEDs and other threat items is not difficult per se, but rather depending on the training of screeners.

A recent study by Koller et al. (2007) showed that training not only has an effect on the detection of trained object views – which then are available in visual memory – but also generalizes to similar looking but untrained views of other objects. This transfer effect was revealed for all threat categories, i.e. for guns, IEDs and other threat items. For knives no transfer effect was found. This could be due to the shape of the knives. On one hand, knives show less diagnostic features which play an important role in object recognition compared to objects from other categories. These few diagnostic features might also get lost when a knife is rotated. On the other hand, the visual similarity of knives to harmless everyday objects (e.g., pen) is substantial. These factors could impede detectability and trainability of knives and ultimately might have resulted in rather small training and transfer effects for knives. They might also be an explanation for the fact that for knives there is the largest viewpoint effect compared to objects from other categories (see Figure 18). While objects of other threat categories in a rotated view usually still show many diagnostic features and also, due to their larger surface, more information in general, for knives much information might be lost with a high rotation angle. Therefore, the detection and the discrimination of knives and harmless objects is hindered. However, rather than being a category generalisation of the gained knowledge, the transfer effect could also have resulted because of a large similarity of the object pairs. This would imply a specific-token familiarity to be the reason for the transfer as Smith, Redford, Gent & Washburn (2005) suggest. If a specific-token familiarity would apply to the recognition of objects then learned knowledge about an object could not be transferred to an object of the same category but only to objects with the same specific tokens. For a more detailed discussion of this issue see Koller et al. (2007) and Smith et al. (2005).

Overall, these results are fully consistent with earlier results (Schwaninger, 2004; Hardmeier et al., 2006; Schwaninger & Hofer, 2004; Schwaninger,

2003) and show that adaptive CBT such as X-Ray Tutor can be a powerful tool to increase efficiency in x-ray image interpretation by airport security screeners.

4.2.6 Acknowledgements

This research was financially supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403) and the Ministry of Justice, The Netherlands.

4.2.7 References

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*, New York: Erlbaum, Hillsdale.

Graf, M., Schwaninger, A., Wallraven, C. & Bülthoff, H.H. (2002). ”Psychophysical results from experiments on recognition & categorisation,” *Information Society Technologies (IST) programme, Cognitive Vision Systems – CogVis (IST-2000-29375).*

Green, D.M., & Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*, New York: Wiley.

Hardmeier, D., Hofer, F. & Schwaninger, A. (2006). “Increased detection performance in airport security screening using the x-ray ort as pre-employment assessment tool,” in *Proceedings of the 2nd International Conference on Research in Air Transportation, ICRAT 2006, Belgrade, Serbia and Montenegro,* June 24-28, 2006, pp 393-397.

Hardmeier, D., Hofer, F. & Schwaninger, A. (2006). “The role of recurrent CBT for increasing aviation security screeners’ visual knowledge and abilities needed in x-ray screening,” in *Proceedings of the 4th International Aviation Security Technology Symposium, Washington, D.C., USA,* November 27 – December 1, pp 338-342.

Hardmeier, D., Hofer, F., & Schwaninger, A. (2005). “The x-ray object recognition test (x-ray ort) – a reliable and valid instrument for measuring visual abilities needed in x-ray screening,” in *IEEE ICCST Proceedings,* vol 39, 2005, pp 189-192.

Hofer, F. & Schwaninger, A. (2004). “Reliable and valid measures of threat detection performance in X-ray screening,” in *IEEE ICCST Proceedings*, vol 38, pp 303-308.

Hofer, F., Hardmeier, D. & Schwaninger, A. (2006). “Increasing airport security using the X-ray ORT as effective pre-employment assessment tool,” in *Proceedings of the 4th International Aviation Security Technology Symposium, Washington, D.C., USA,* November 27 – December 1, 2006, pp 303-308.

Koller, S. & Schwaninger, A. “Assessing X-ray image interpretation competency using the X-Ray Tutor Competency Assessment Test X-Ray CAT,” in *Proceedings of the 2nd International Conference on Research in Air Transportation, ICRAT 2006, Belgrade, Serbia and Montenegro,* June 24-28, 2006, pp 399-402.

Koller, S.M., Hardmeier, D. Michel, S. & Schwaninger, A. (2007). “Investigating training and transfer effects resulting from recurrent CBT of x-ray image interpretation,” in D.S. McNamara and J.G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society,* 2007, pp 1181-1186.

MacMillan, N.A. & Creelman, C.D. (1991). *Detection theory: A user’s guide*, Cambridge: University Press.

McCarley, J.S., Kramer, A.F., Wickens, C.D., Vidoni, E.D. & Boot, W.R. (2004). “Visual skills in airport screening, *Psychological Science*, vol 15(5), pp 302-306, 2004.

Schwaninger, A. (2003). “Evaluation and selection of airport security screeners,” *AIRPORT*, vol 02/2003, pp 14-15.

Schwaninger, A. (2003). “Training of airport security screeners,” *AIRPORT*, vol 05/2003, pp 11-13.

Schwaninger, A. (2004). “Computer based training: a powerful tool to the enhancement of human factors,” *Aviation Security International*, vol FEB/2004, pp 31-36.

Schwaninger, A. (2004). “Objekterkennung und Signaldetektion,“ In: B. Kersten & M.T. Groner (Eds.), *Praxisfelder der Wahrnehmungspsychologie* Bern: Huber, pp 106-130.

Schwaninger, A. & Hofer, F. (2004). “Evaluation of CBT for increasing threat detection performance in X-ray screening,” In: K. Morgan and M. J. Spector, *The Internet*

Society 2004, Advances in Learning, Commerce and Security, Wessex: WIT Press, pp 147-156.

Schwaninger, A., Hardmeier, D., & Hofer, F. (2005). “Aviation security screeners visual abilities & visual knowledge measurement,” *IEEE Aerospace and Electronic Systems*, vol 20(6), pp 29-35.

Schwaninger, A.,Bridges, A.P., Drury, C., Durinckx, F., Durrant, P., Hodge, T., Hofer, F., Jongejan, R., Maguire, R.L., McClumpha, A., Neiderman, E., Steinmann, C. & W. Wüest (2005). *Principles and Requirements for Assessing X-Ray Image Interpretation Competency of Aviation Security Screeners,* White Paper, International Aviation Security Human Factors Technical Advisory Group (InterTAG), Competency Assessment Working Group (CAWG).

Smith, J.D., Redford, J.S., Gent, L.C. & Washburn, D.A. (2005). “Visual search and the collapse of categorization,” *Journal of Experimental Psychology: General*, vol 134(4), pp 443-460.

4.3 Investigating training, transfer, and viewpoint effects resulting from recurrent CBT of x-ray image interpretation

4.3.1 Abstract

X-ray screening of passenger bags is an essential task at airport security checkpoints. In this study we investigated how well airport security screeners can detect guns, knives, improvised explosive devices (IEDs) and other threat objects in x-ray images of passenger bags before and after three and six months of recurrent (about 20 min per week) computer-based training (CBT). Two experiments conducted at different airports gave very similar results. Training with X-Ray Tutor (XRT), an individually adaptive CBT, resulted in large performance increases, especially for detecting IEDs. While performance for detecting IEDs was initially substantially lower than for guns, IEDs could be detected as well as guns after several months of training. A large transfer effect was observed as well: Training with XRT helped screeners recognize new threat objects that were similar in shape as the trained objects. Threat recognition was dependent on the rotation of the objects. If depicted from an unusual viewpoint, prohibited items were more difficult to recognize. The results were compared to two conventional (not adaptive) CBT systems. For one system no training and transfer effects were observed whereas small training and transfer effects were found for the other conventional CBT system.

4.3.2 Introduction

The importance of aviation security has increased dramatically in the last years. As a consequence of the new threat situation, large investments were made into modern security technology. State of the art x-ray screening equipment offers good image quality, high resolution and many image enhancement functions. However, the decision whether an x-ray image of a passenger bag contains a prohibited item or not, is still being taken by a human operator, i.e. an airport security screener. Object shapes that are not similar to ones stored in visual memory are difficult to recognize (e.g., Graf, Schwaninger, Wallraven, and Bülthoff, 2002; Schwaninger, 2004, 2005). Schwaninger, Hardmeier, and Hofer (2005) have shown that x-ray screener performance depends on knowledge-based and image-based factors. A prerequisite for good x-ray detection performance is knowledge about which objects are prohibited and what they look like in x-ray images. Such knowledge is acquired by computer-based, class-room and on the job training (knowledge-based factors). Image-based factors refer to image difficulty resulting from viewpoint variation of threat objects, superposition of threat objects by other objects in a bag, and bag complexity depending on the number and type of other objects in the bag. The ability to cope with image-based factors is related to individual visual-cognitive abilities rather than a mere result of training (Hardmeier, Hofer, and Schwaninger, 2006).

Computer-based training is expected to be a very important determinant of x-ray image interpretation competency, because many threat objects are not known from everyday experience and because objects look quite different in x-ray images than in reality. This is illustrated in Figure 20 with two examples.

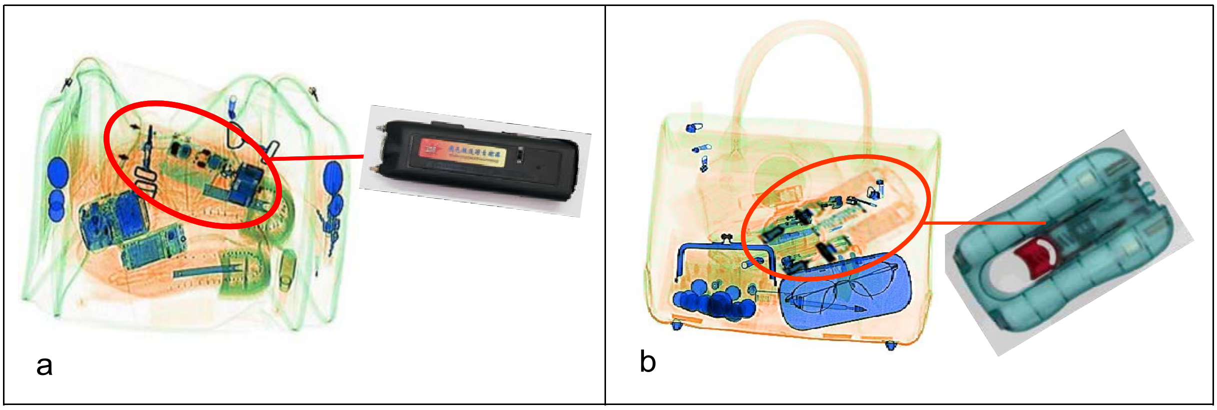

Figure 20: Different types of prohibited items in x-ray images of passenger bags. a) Electric shock device, b) self defense gas spray “Guardian Angel”.

Schwaninger and Hofer (2004) and Schwaninger, Wetter and Hofer (2007) could show that detection of improvised explosive devices (IEDs) in hold baggage screening (HBS) can be significantly improved if people are trained with an individually adaptive training system such as X-Ray Tutor (XRT). Schwaninger et al. (2005) compared detection performance of novices with the one of aviation security screeners. A rather poor recognition of unfamiliar object shapes (e.g., self-defense gas spray, electric shock device etc.) in x-ray images was found for novices. For experienced aviation security personnel, a much higher recognition performance was observed. McCarley, Kramer, Wickens, Vidoni and Boot (2004) reported a better performance after training for the detection of knives in x-ray images for novices.

When one takes into account the myriad of views that can be produced by a single object, the question arises how the human brain stores and recognizes objects even if they are presented in unusual views. In the object recognition literature, two types of theories can be distinguished: structural description theories and view-based theories. The former assume that objects are stored in visual memory by their component parts and their spatial relationship. An object-centered description of this nature was described by Marr and Nishihara (1978), who proposed that objects are hierarchically decomposed into their parts and spatial relations relative to object-centered coordinates in order to access an object-centered 3D model in visual memory. In Biederman´s (1987) recognition by components (RBC) theory, non-accidental properties like vertices, parallel vs. non-parallel lines, straight vs. curved lines etc. (see Lowe, 1985, 1987) are extracted from a line drawing representation of objects to define basic geometrical primitives (geometrical ions, "geons") that are relatively orientation-invariant. A geon structural description (GSD) in memory is activated by extracting geons from the visual input and match geon properties and their spatial relationship with the GSD (Hummel and Biederman, 1992).

For view-based theories, different approaches have been proposed. Examples are recognition by alignment to a 3D representation (Lowe, 1987), recognition by linear combination of 2D views (Ullman and Basri, 1991), recognition by view interpolation (e.g., using RBF networks) proposed by Poggio and Edelman (1990) and storing of multiple views for each object plus performing transformations (Tarr and Pinker, 1989). What view-based theories have in common is the assumption that objects are not stored in memory as rotation invariant structural descriptions but instead in a format which is viewer-centered. A more detailed discussion of structural description theories vs. view-based theories and more recent hybrid theories is beyond the scope of this paper (for reviews see for example Graf, Schwaninger, Wallraven and Bülthoff, 2002; Hayward, 2003; Kosslyn, 1994; Peissig and Tarr, 2007; Schwaninger, 2005; Tarr and Bülthoff, 1998). However, it should be pointed out that empirical results seem to be correlated with the required level of recognition (Bülthoff, Edelman and Tarr, 1995; Tarr, 1995): if the object has to be recognized at 'entry level', behavioral measures are less affected by changes in perspective. However, in the case of subordinate recognition in which fine discriminations are typically required, both response times and accuracy are more sensitive to the specific viewpoint used. Furthermore, differences in the task a subject has to perform (Lawson, 1999) and the specific paradigm that is used (Verfaillie, 1992) can influence which level of representation is tapped (see also Logothetis and Sheinberg, 1996).

The first aim of this study is to investigate how well airport security screeners can detect guns, knives, IEDs and other prohibited items in x-ray images of passenger bags. The second aim is to examine whether screener detection performance can be increased by conducting recurrent CBT. To this end, screeners conducted weekly recurrent CBT (about 20 min per week). Detection performance was tested with the X-Ray Competency Assessment Test (X-Ray CAT) by Koller and Schwaninger (2006). This test measures how well people detect threat items in x-ray images of passenger bags. It was conducted at the beginning and then after three and six months of training. In addition to training effects, the X-Ray CAT allows measuring transfer effects, i.e. to what extent visual knowledge that was gained through CBT can be transferred to other threat items (see below). In the X-Ray CAT all prohibited items are depicted from a canonical (easy recognizable) perspective (Palmer, Rosch and Chase, 1981) and unusual perspective which allows investigating viewpoint effects. The study was conducted at two mid-size European airports. In Airport 1 (Experiment 1) one group of screeners used adaptive CBT (XRT) whereas the other group of screeners (control group) used a conventional (not adaptive) CBT. In

Airport 2 (Experiment 2) the same experimental design was used except for the fact that the control group used another conventional CBT system. This allows investigating whether a training effect is dependent on the type of the CBT system used.

4.3.3 Experiment 1

4.3.3.1 Method

4.3.3.1.1 Participants

A total of 209 airport security screeners of a mid-size European airport participated in Experiment 1 and conducted the X-Ray CAT 1.0.0 three times with an interval of three months between the measurements. The adaptive CBT group (XRT group) consisted of 97 screeners who conducted weekly recurrent CBT using X-Ray Tutor (XRT) CBS 2.0 Standard Edition between all three test measurements. The control group consisted of 112 screeners who used a conventional (not adaptive) CBT. According to the security organization and their Appropriate Authority, airport security screeners of both groups conducted about 20 min CBT per week. Analysis of XRT training use showed that on average, each screener trained 20.26 minutes (*SD* = 3.65 min) per week.

4.3.3.2 Materials and Procedure

4.3.3.2.1 X-Ray Competency Assessment Test (X-Ray CAT)

The X-Ray CAT consists of 256 trials based on 128 different color x-ray images of passenger bags. Each of the bag images is used once containing a prohibited item (threat image) and once without any threat object (non threat image). Figure 21 displays examples of the stimuli. Note that in the test the images are displayed in color.

Prohibited objects can be assigned to four categories as defined in Doc 30 of the European Civil Aviation Conference (ECAC): guns, IEDs, knives and other prohibited items (e.g., self-defense gas spray, chemicals, grenades etc.). The threat objects have been selected and prepared in collaboration with experts of Zurich State Police, Airport Division to be representative and realistic. For each threat category 16 exemplars are used (eight pairs). Each pair consists of two prohibited items that are similar in shape (see Figure 22). These were distributed randomly into two sets, set A and set B. Prohibited items of set A (non threat bag images) are contained in the XRT CBS 2.x SE training whereas the items of set B are not. This allows testing for transfer effects. Every item is depicted from two different viewpoints. The easy viewpoint refers to the canonical (i.e. easy recognizable) perspective (Palmer et al., 1981). The difficult viewpoint shows the threat item with an 85 degree horizontal rotation or an 85 degree vertical rotation relative to the canonical view (see Figure 22 for examples). In each threat category, half of the prohibited items of the difficult viewpoint are rotated vertically, the

Figure 21: Example images from the X-Ray CAT. Left: harmless bag (non threat image), right: same bag with a prohibited item at the top right corner (threat image). The prohibited item (gun) is shown also separately at the bottom right.

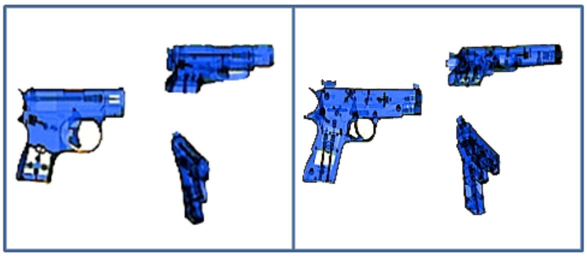

Figure 22: Example of two x-ray images of similar looking threat objects used in the test. Left: a gun of set A. Right: Corresponding gun of set B.

other half horizontally. Set A and B are equalized concerning the rotations of the prohibited objects.

Every threat item is combined with a bag in a manner that the degree of superposition by other objects is similar for both viewpoints. This was achieved using a function that calculates the difference between the pixel

$$SP = \frac{\sqrt{\sum [I_{SN}(x,y) - I_N(x,y)]^2}}{ObjectSize}$$

SP = Superposition; I_{SN} = Grayscale intensity of the SN (Signal plus Noise) image (contains a prohibited item); I_N = Grayscale intensity of the N (Noise) image (contains no prohibited item); Object Size: Number of pixels of the prohibited item where R, G and B are < 253

intensity values of the bag image with the threat object minus the bag image without the threat object using the following formula:

Using this equation (division by object size), the superposition value is independent on the size of the prohibited item. This value can be kept relatively constant for the two views of a threat object, independent of the degree of clutter in a bag, when combining the bag image and the prohibited item. The bag images were visually inspected by aviation security experts to ensure they do not contain any other prohibited items. Harmless bags were assigned to the different categories and viewpoints of the threat objects in a way that their difficulty was balanced across all categories[3]. The false alarm rate (the rate at which screeners wrongly judged a harmless bag as containing a threat item) for each bag image served as measure of difficulty based on a pilot study with 192 screeners of another airport.

[3] The eight categories of test images (four threat categories in two viewpoints each) are similar in terms of the difficulty of the harmless bags. This means, a difference of detection performance between categories or viewpoints can not be due to differences in the difficulty of the bag images.

The X-Ray CAT takes about 30-40 minutes to complete. Each image is shown for a maximum of 10 seconds on the screen. Screeners have to judge whether the bag is OK (contains no prohibited item) or NOT OK (contains a prohibited item). Additionally, screeners have to indicate the perceived difficulty of each image on a 100 point scale (difficulty rating)[4]. The X-Ray CAT is built into the XRT training system (see below). The interface of the X-Ray CAT is the same as in XRT except there is no feedback and screeners do not have to click on the image to identify the threat object.

4.3.4 X-Ray Tutor (XRT) Training System

X-Ray Tutor (XRT) is an individually adaptive training system for aviation

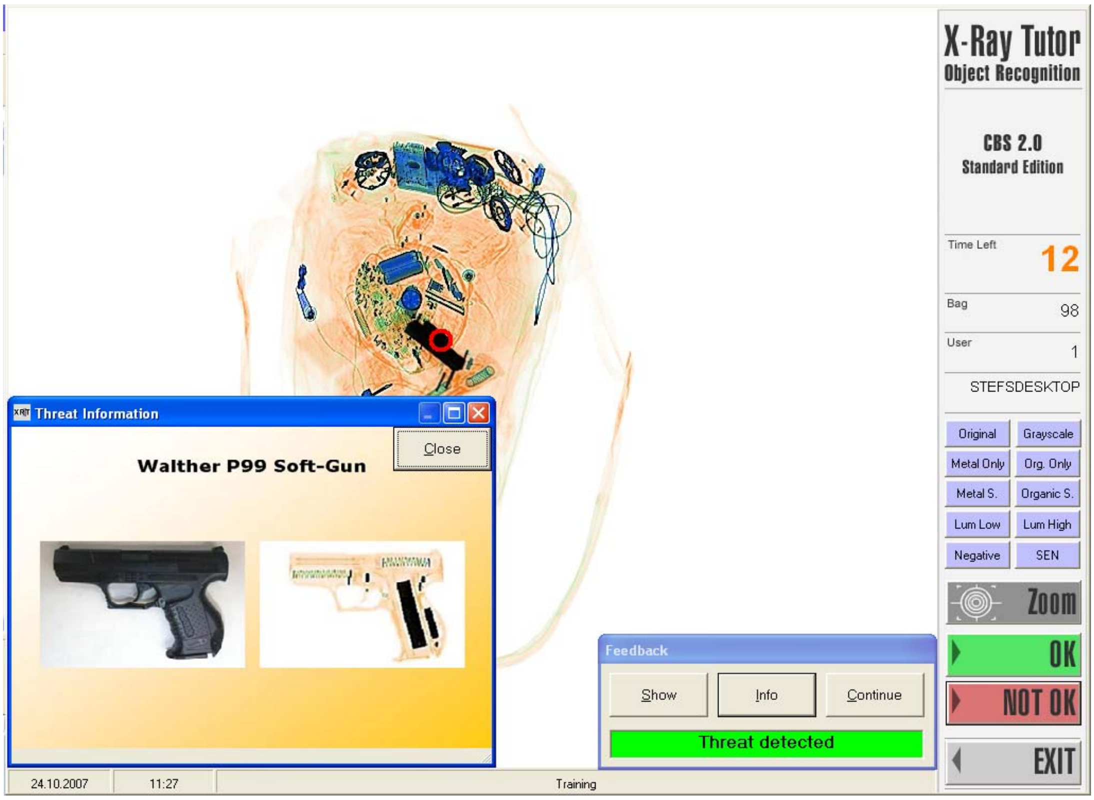

Figure 23: Screenshot of the XRT CBS 2.0 training system during training. At the bottom right a feedback is provided after each response. If a bag contains a prohibited item, an information window can be displayed (see bottom left of the screen).

[4] The difficulty ratings were not analyzed in this study.

security screeners. It contains a large image library with hundreds of different threat objects depicted in up to 72 views, more than 6000 bag images and many millions of possible threat object to bag combinations (see Schwaninger, 2004 for details). The individually adaptive training algorithm of XRT starts with showing threat objects depicted from easy viewpoints with little superposition by other objects and in bags of low complexity. Based on each individual screeners´ learning progress, threat objects are shown in more difficult views, more complex bags and with more superposition. These parameters are adapted automatically by a scientifically validated algorithm for each screener and threat object while taking into account automatic image processing algorithms as explained in Schwaninger, Michel and Bolfing (2007). XRT first presents screeners prohibited objects in easy (canonical) views. The individually adaptive training algorithm determines for each screener which views are difficult to recognize and adapts the training so that the trainee becomes able to detect threat items reliably even if prohibited objects are substantially rotated away from the easiest view. During the next difficulty levels, first superposition and then bag complexity is increased so that the trainee becomes able to detect threat items reliably even if they are superimposed by other objects or if the complexity of a bag is very high (for more information on XRT see Schwaninger, 2003, 2004, 2005b).

During a training session each image is displayed for 15 seconds on the screen. Within this time screeners can use image enhancement functions which are also available when working with the x-ray machine (e.g., grayscale, negative image, edge enhancement, etc.). If the image contains a prohibited item, screeners have to click on it and then click on the NOT OK button. If the bag is harmless; they have to click on the OK button. After providing a confidence rating using a slider control, feedback is shown to inform the trainee whether the image has been judged correctly or not (see Figure 23). If the bag contains a threat item, it is highlighted by flickering and the trainee has the possibility to display information about the threat item (see bottom left of Figure 23). By clicking on the continue button the next image is shown. As a default setting, one training sessions takes 20 minutes. During this time screeners see between 150 and 300 images.

4.3.5 Procedure

As explained above, two groups of screeners participated in Experiment 1. The XRT training group conducted weekly recurrent CBT using XRT CBS 2.0 Standard Edition. The control group used a conventional (not adaptive) CBT. In order to avoid potential negative consequences, we decided not to mention the exact CBT product in this article. However, it can be mentioned that this CBT is also widely used at many airports worldwide. It has a much smaller threat image library than XRT, threat objects are not displayed in many different views, threat objects are not matched with different bags on the fly, and there is no individually adaptive training algorithm.

The XRT training group and the control group took the X-Ray CAT before, after three, and after six months of weekly CBT. This allows testing the effectiveness of both CBT systems for increasing x-ray image interpretation competency of airport security screeners. As explained above, half of the prohibited items in the X-Ray CAT are also contained in the XRT training system (although presented in different bags). The other half of the prohibited items of the X-Ray CAT is not part of the XRT training library. This allows testing for transfer effects, i.e. testing whether training with the detection of certain prohibited items helps increasing the detection of other prohibited items. Finally, as specified above in the section on the X-Ray

CAT, all prohibited items are depicted in easy and difficult view which allows testing effects of viewpoint on screener detection performance.

4.3.6 Results and Discussion

Effect size	d	η^2
small	0.20-0.49	0.01-0.05
medium	0.50-0.79	0.06-0.13
large	≥ 0.8	≥ 0.14

Table 8: Classification of effect sizes based on Cohen (1988)

Detection performance was calculated using the signal detection measure d' (Green and Swets, 1966), which takes into account the hit rate (correctly judged threat images as being NOT OK) and the false alarm rate (wrongly judged harmless bags as being NOT OK). D' is calculated using the following formula: d´= z(H) – z(FA). Whereas H is the hit rate, FA the false alarm rate and z refers to the z-transformation. Performance values are not reported due to security reasons. However, effect sizes are reported for all relevant analyses and interpreted based on Cohen (1988), see Table 8. For *t*-tests, *d* between 0.20 and 0.49 represents small effect size; *d* between 0.50 and 0.79 represents medium effect size; *d* ≥ 0. 80 represent large effect size. For analysis of variance (ANOVA) statistics, η^2 between 0.01 and 0.05 represents small effect size; η^2 between 0.06 and 0.13 represents medium effect size; $\eta^2 \geq 0.14$ represent large effect size.

Figure 24 shows the detection performance of the first, second and third measurement for both screener groups. As can be seen in the Figure, there was a large improvement as a result of training in the XRT training group while there was no improvement in the control group. These results were confirmed by an ANOVA for repeated measures using d' scores with the within-participant factor measurement (first, second and third) and the

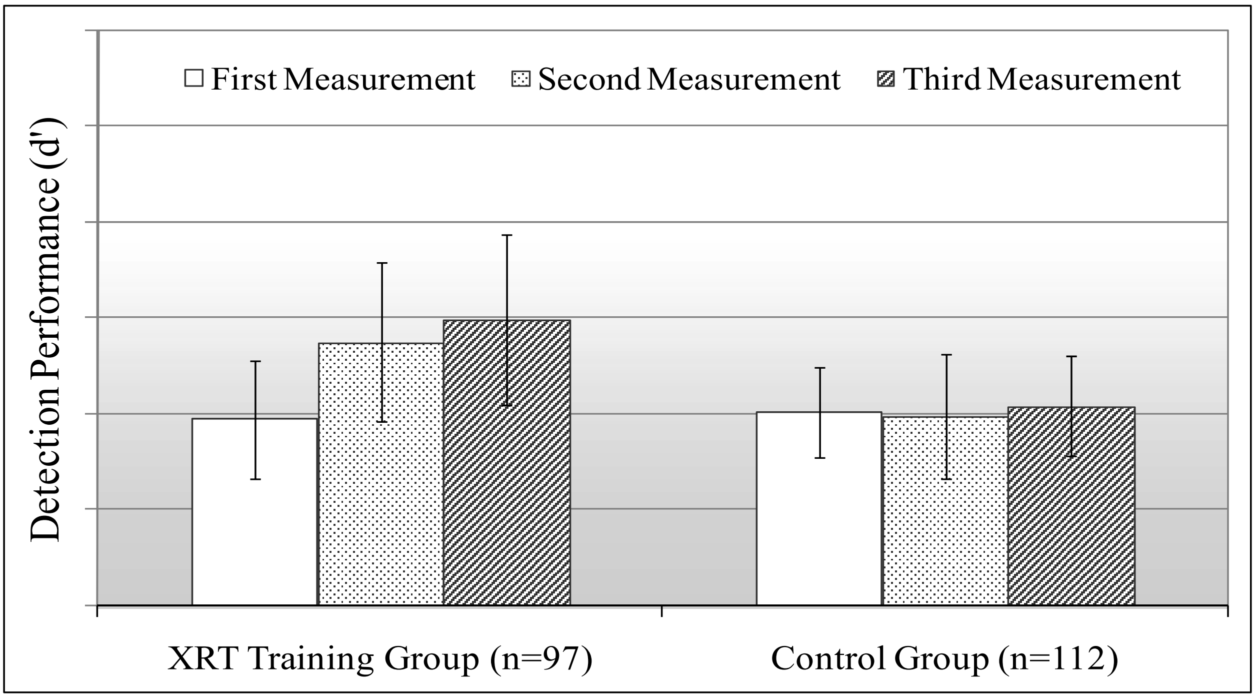

Figure 24: Detection performance with standard deviations for the XRT training group (left) vs. the control group (right) comparing first, second and third measurement

between-participant factor group (XRT training group and control group). There were large main effects of measurement, $\eta^2 = .28$, $F(2, 414) = 81.04$, $p < .001$, and group, $\eta^2 = .19$, $F(1, 207) = 47.62$, $p < .001$. There was also a large interaction of measurement and group, $\eta^2 = .25$, $F(2, 414) = 68.67$, $p < .001$, which is consistent with Figure 24 showing large performance increases as a result of training only for the XRT training group but not for the control group.

Separate pairwise *t*-tests of detection performance d' revealed no significant difference at the baseline measurement between the two groups $t(177) = -0.91$, $p = .363$, $d = 0.13$, but already a significant difference in the second measurement, i.e. after three months of training, $t(207) = 7.52$, $p < .001$, $d = 1.04$. Additional paired-samples *t*-tests revealed significant differences for the XRT training group between all three test measurements but no significant differences for the control group (see Table 9).

Table 9: Results of the *t*-tests comparing the detection performance of first (t1), second (t2) and third (t3) measurement

	***t*(96)**	***p***	***d***
XRT Training Group (t1 – t2)	-9.80	< .001	1.12
XRT Training Group (t2 – t3)	-3.95	< .001	0.28
	***t*(111)**	***P***	***d***
Control Group (t1 – t2)	.54	= .59	0.05
Control Group (t2 – t3)	-1.89	= .06	0.17

Figure 25 shows the detection performance of both screener groups broken up by prohibited item category and the three test measurements. A repeated-measures ANOVA with the within-participant factors

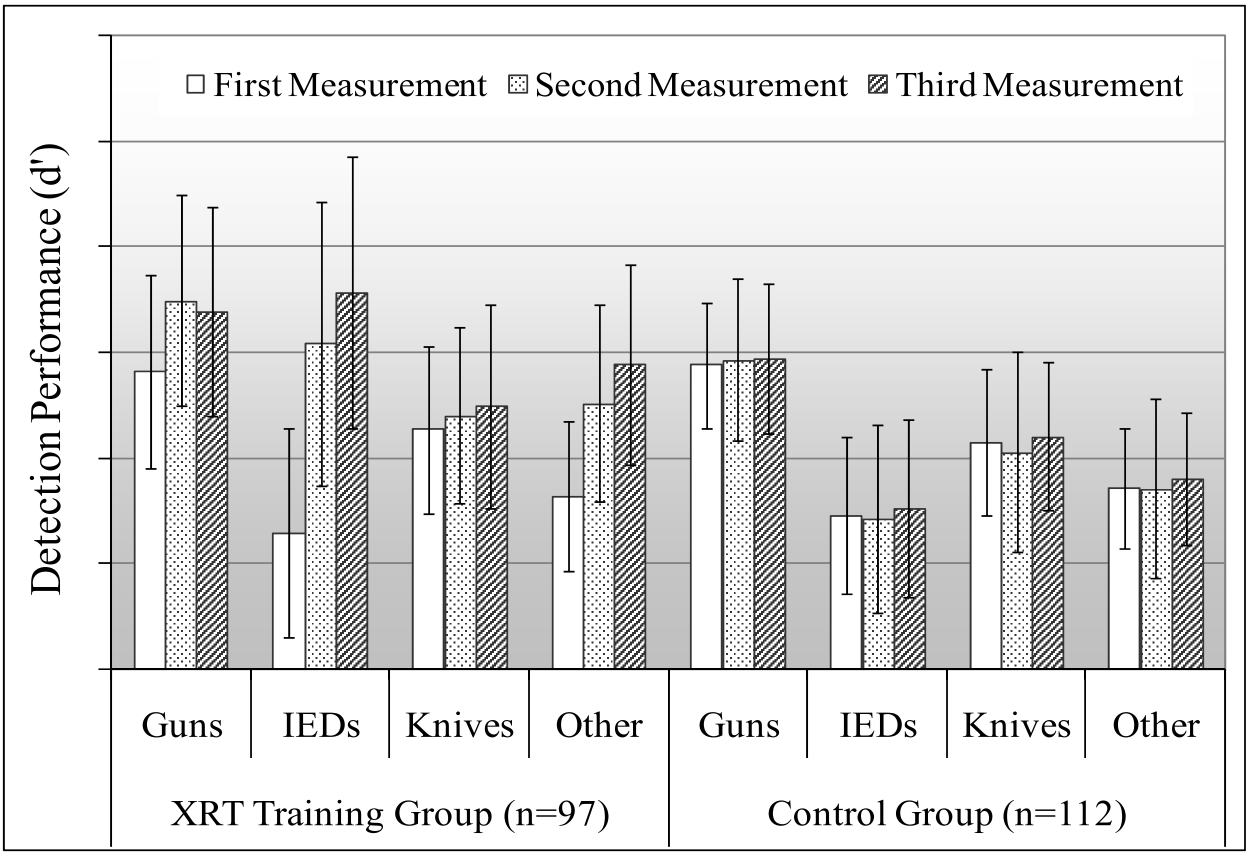

Figure 25: Detection performance with standard deviations for the XRT training group vs. the control group broken up by prohibited item category and test measurement.

measurement (first, second and third) and threat category (guns, IEDs, knives and other), and the between-participant factor group (XRT training vs. control) revealed the significant main effects and significant interactions given in Table 10a. In addition to the effects that were already found in the previous ANOVA, also the factor threat (or prohibited item) category was significant. As can be seen in Figure 25, guns were detected best, followed by knives, other prohibited items and IEDs at the first test measurement. There was a highly significant interaction between threat category and measurement. As can be seen in Figure 25, detection of IEDs was initially much lower than gun detection. After six months of training, screeners of the XRT training group could detect IEDs even slightly better than guns. This result implies that IED detection is not difficult per se but rather a matter of the right training. Note that in this study all IEDs contained a detonator, wires, explosive, a triggering device and a power source. Thus our conclusions are only applicable to the detection of such multi-component IEDs. Large performance increases were also found for other prohibited items in this group, while for knives, only a small improvement as a result of training was found. Note that after six months of training, detection performance of knives is lower than the one for any other threat category in the XRT training group; although at baseline measurement it was higher than the detection performance for IEDs or other threat objects. The interaction between threat category, group and measurement is also worth mentioning. As can be seen in Figure 25 this results from the fact that there was no training effect for the control group. Their detection performance remains at about the same level for each threat category even after six months of training with the conventional (not adaptive) CBT system.

Separate pairwise *t*-tests were conducted to compare detection performance at the first and the second measurement for both groups and each threat category separately (Table 11). The XRT training group showed a significant increase of the detection performance at the second measurement for the categories guns, IEDs and other threat objects. For knives, a significant difference could be found only in the third measurement. The comparison of the effect size *d* between the *t*-tests of the four threat categories confirms the earlier mentioned conclusion that the training effect was particularly big for IEDs and rather small for knives. Detection performance of the control group did not differ significantly between the measurements, confirming that the conventional CBT did not result in an increase of threat detection performance.

Table 10: Results of the ANOVAs in Experiment 1

	Factor	**df**	**F**	**η^2**	***p***
a)	Measurement (M)	2, 414	83.96	.29	< .001
	Threat Category (T)	3, 621	240.03	.54	< .001
	Group (G)	1, 207	56.20	.21	< .001
	M x G	2, 414	70.49	.25	< .001
	T x G	3, 621	45.05	.18	< .001
	M x T	6, 1242	43.20	.17	< .001
	M x T x G	6, 1242	40.65	.16	< .001
b)	Measurement (M)	2, 414	80.55	.28	< .001
	Set (S)	1, 207	4.18	.02	< .05
	Group (G)	1, 207	49.40	.19	< .001
	M x G	2, 414	67.99	.25	< .001
	M x S	2, 414	8.80	.04	< .001
	S x G	1, 207	51.32	.20	< .001
	M x S x G	2, 414	11.54	.05	< .001
c)	Measurement (M)	2, 414	87.69	.30	< .001

	Set (S)	1, 207	2.37	.01	= .13
	Threat Category (T)	3, 621	236.79	.53	< .001
	Group (G)	1, 207	63.57	.24	< .001
	M x G	2,414	71.16	.26	< .001
	M x T	6, 1242	44.35	.18	< .001
	M x S	2, 414	10.93	.05	< .001
	S x G	1, 207	52.25	.20	< .001
	S x T	3, 621	74.00	.26	< .001
	T x G	3, 621	47.39	.19	< .001
	M x T x G	6, 1242	41.04	.17	< .001
	M x S x G	2, 414	10.74	.05	< .001
	M x S x T	6, 1242	3.84	.02	< .01
	S x T x G	3, 621	4.78	.02	< .01
	M x S x T x G	6, 1242	2.99	.01	< .01
d)	Measurement (M)	2,414	84.10	.29	< .001
	View (V)	1, 207	1768.63	.90	< .001
	Threat Category (T)	3, 621	258.62	.56	< .001
	Group (G)	1, 207	61.91	.23	< .001
	M x G	2, 414	65.80	.24	< .001
	M x T	6, 1242	41.33	.17	< .001
	M x V	2, 414	2.05	.01	= .13

	V x G	1, 207	3.27	.02	= .07
	V x T	3, 621	425.64	.67	< .001
	T x G	3, 621	40.86	.17	< .001
	M x T x G	6, 1242	40.25	.16	< .001
	M x V x G	2, 414	2.23	.01	< .05
	M x V x T	6, 1242	6.58	.03	< .001
	V x T x G	3, 621	3.08	.02	< .05
	M x V x T x G	6, 1242	2.68	.01	< .05

Table 11: Results of the *t*-tests comparing the detection performance of the four categories between the first (t1), second (t2) and third (t3) measurement

XRT training group	***t*(96)**	**df**	***p***	***d***
Guns t1 – t2	- 5.96	96	< .001	0.70
IEDs t1 – t2	- 13.03	96	< .001	1.53
Knives t1 – t2	- 1.51	96	= .13	0.17
Other t1 – t2	- 8.47	96	< .001	1.07
Guns t1 – t3	- 4.69	96	< .001	0.60
IEDs t1 – t3	- 15.88	96	< .001	2.00
Knives t1 – t3	- 2.27	96	< .05	0.26

Other t1 – t3	- 12.56	96	< .001	1.51
Control group	***t*(111)**	**df**	***p***	***d***
Guns t1 – t2	- 0.40	111	= .69	0.05
IEDs t1 – t2	0.03	111	= .98	0.00
Knives t1 – t2	0.83	111	= .41	0.09
Other t1 – t2	-0.17	111	= .87	0.02
Guns t1 – t3	-0.92	111	= .36	0.10
IEDs t1 – t3	-1.05	111	= .30	0.08
Knives t1 – t3	-0.73	111	= .47	0.08
Other t1 – t3	-1.39	111	= .17	0.15

The results of the analyses considering the two prohibited item sets of the X-Ray CAT, set A and set B, are shown in Figures 26 and 27. As explained

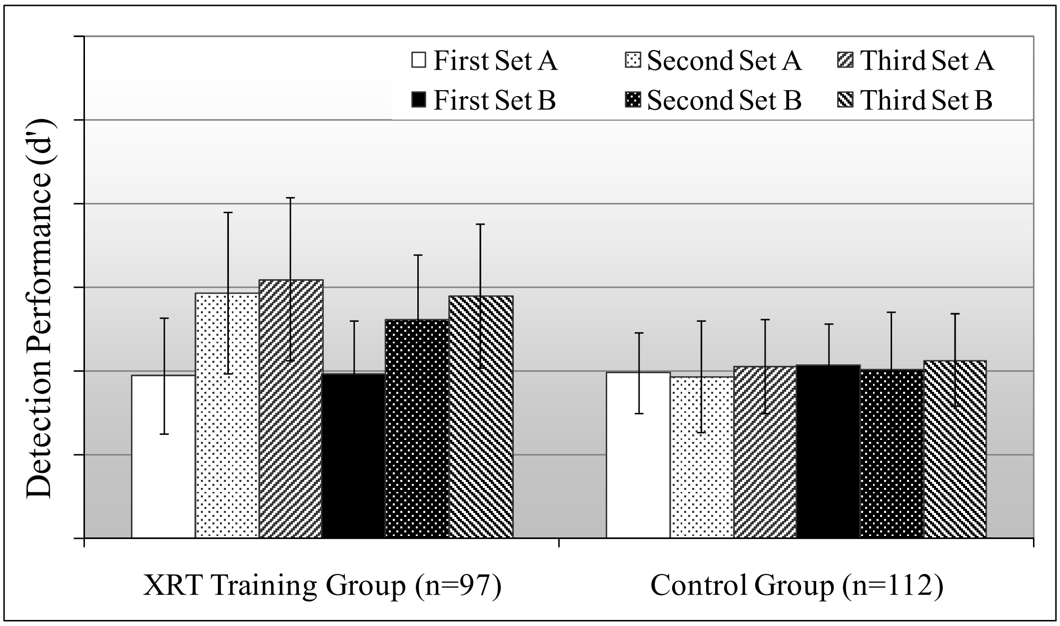

Figure 26: Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for set A and set B separately

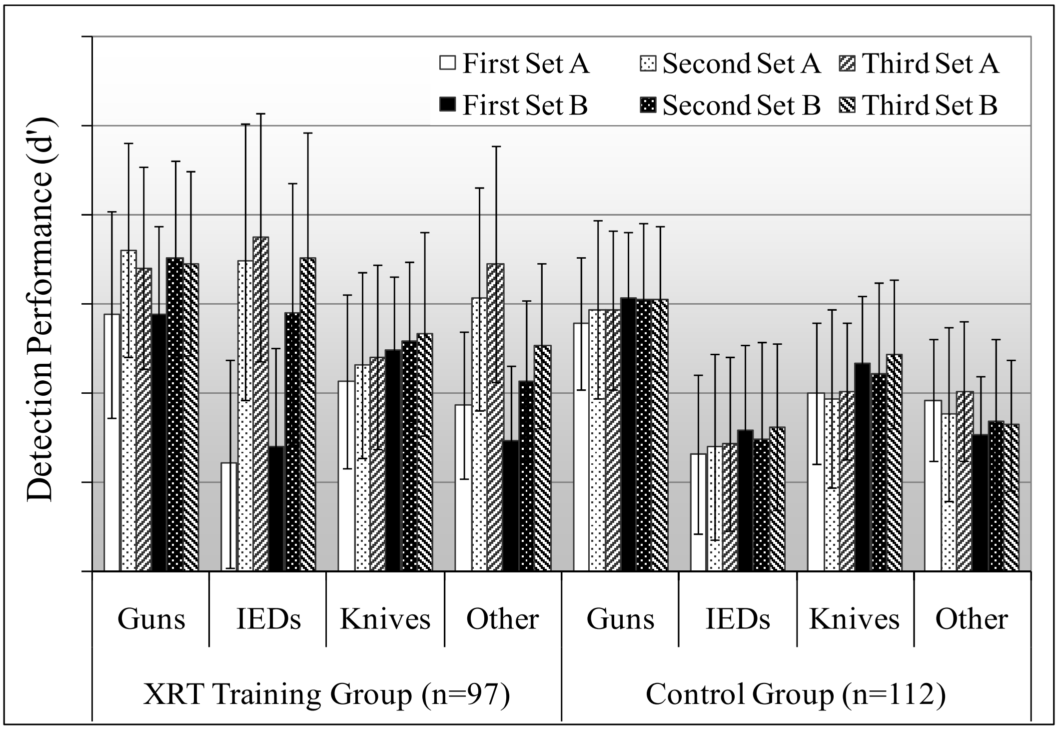

Figure 27: Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for set A and set B and each threat category separately

above, set A items are X-Ray CAT images which contain prohibited items which are part of the XRT image library. Set B items are X-Ray CAT images which contain prohibited items that are not part of the XRT image library. By comparing training effects for set A and set B transfer effects can be investigated, i.e. whether training with XRT does not only improve detection of prohibited items that are part of the XRT image library (set A) but also the detection of other prohibited items that are visually similar (set B). Figure 27 shows the detection performance for both screener groups broken up by test set for all three measurements. It shows a clear increase in detection performance for the XRT training group, especially at the second measurement, after the first three months of training. For the control group, as in the previous analysis, no training effect is evident. The results of the repeated measures ANOVA with the within-participant factors measurement (first, second and third) and set (A vs. B) and the between-participant factor group (XRT training group vs. control group) can be seen in Table 10b. There was a significant effect of set in this analysis, which would imply a different detection performance for set A vs. set B. However, the effect is very small, as the effect size of $\eta 2= 0.2$ clearly shows, which makes the difference quasi negligible. This is also supported by the small effect size for the interaction between set and measurement, $\eta 2= 0.4$. Pairwise t-tests showed a significant increase in detection performance at the second measurement for both sets for the XRT training group, set A, $t(96) = -10.27$, $p < .001$, $d = 1.19$, set B, $t(96) = - 7.68$, $p < .001$, $d = 0.92$. These results indicate a large transfer effect, i.e. visual knowledge regarding the visual appearance of the prohibited objects of the XRT image library helped screeners to detect similar looking, but untrained objects in the X-Ray CAT (set B). Consistent with previous analyses, there was no training effect for the control group, neither for set A, $t(111) = .76$, $p = .45$, $d = 0.08$, nor for set B, $t(111) = - 0.28$, $p = .78$, $d = 0.03$. Pairwise t-tests comparing both sets within one group at the first measurement revealed a significant difference of the two sets only for the control group $t(111) = - 2.82$, $p < .01$, $d = 0.17$ but not for the XRT training group, $t(96) = - 0.42$, $p = .68$, $d = 0.03$. However, note that an effect size of $d = 0.17$ is very small which supports the assumption that the two sets are in fact very similar in their difficulty level. Figure 27 includes also the threat category in the analysis. The increase in detection performance for the XRT training group can also be seen in the different threat categories. Pairwise t-tests between the first and second measurement confirmed a significant ($p < .001$, all $d > 0.62$) increase in detection performance for the XRT training group for all threat categories per set except for knives (set A: $p = .12$, $d = 0.19$, set B; $p = .32$, $d = 0.12$). In Figure 27, detection performance in Set A for guns shows a decrease between the second and third measurement. However, this difference was not significant ($p = .13$, $d = 0.17$). For the control group, detection performance between the first and third measurement was compared in order to maximize the chances for finding a significant training effect. Even here, for all categories in each set, the detection between the first and third measurement did not differ significantly (all $p > .12$, $d < 0.18$).The extended ANOVA with the additional within-participant factor threat category revealed the main effects and interactions as specified in Table 10c. The main effect of set was not significant but there were significant interactions with set (see Table 10c). However, as can be seen in Figure 27, these interactions are rather small, which implies large transfer effects.

Figure 28 shows the results of the viewpoint analysis. An ANOVA was conducted on d' scores with the within-participant factors measurement, threat category and viewpoint and the between-participants factor group. It showed significant main effects of measurement, category, viewpoint and group. For details and interactions see Table 10d. The large main effect of viewpoint indicates a higher detection performance for objects in easy

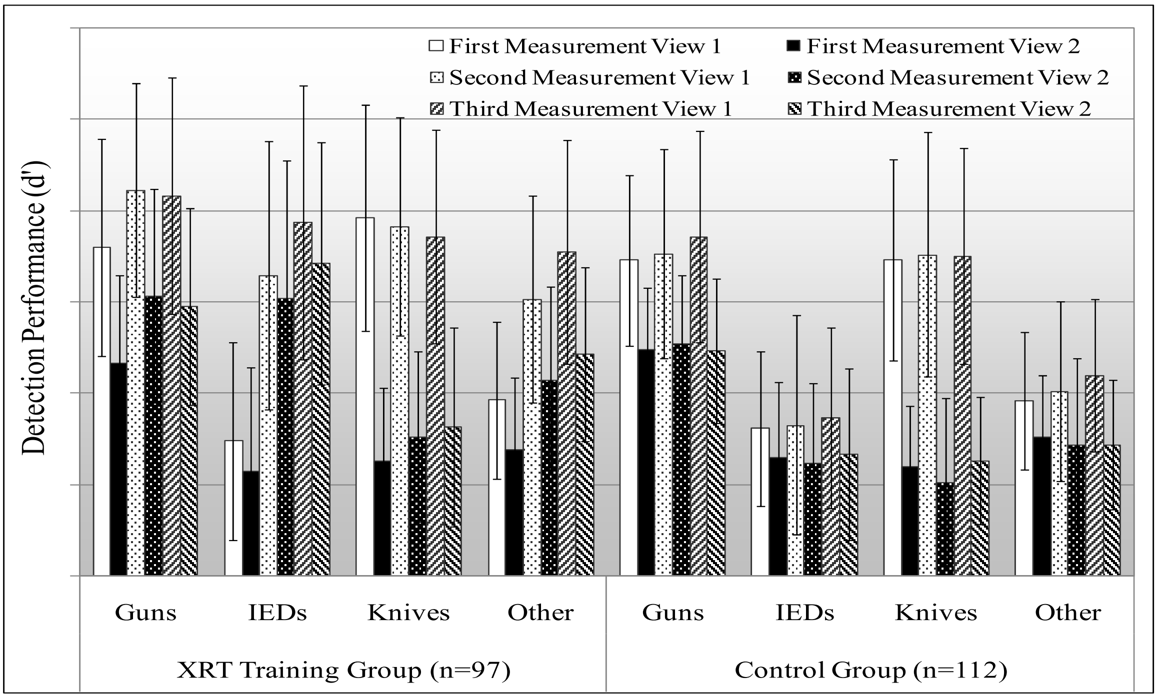

Figure 28: Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for both views and each threat category separately

(canonical) viewpoint compared to objects presented in a difficult (rotated) view (cf. Figure 28).

However, no significant interaction between viewpoint and training could be found. This would suggest that the viewpoint effect is unaffected by the training and could not be decreased. Pairwise *t*-tests showed a significant increase in detection performance at the second measurement for both views in all categories for the XRT training group with the exception of knives in the easy view (p = .53, d = 0.07).

All other comparisons were significant $p < .05$, $d > 0.31$). For the control group no significant increase in detection performance could be found (all $p > .10$, $d < .0.19$), see Table 12 for details. Training with XRT has an effect not only on the objects in the easy view but also on those in the difficult view. The screeners could make the association between the rotated object they detected during training and the canonical view of the object which is displayed in the object information in XRT.

Table 12: Results of the *t*-tests comparing the detection performance of the four categories for easy view (V1) and difficult view (V2) between the first (t1) and second (t2) measurement

XRT training group	***t*(96)**	***p***	***d***
Guns: V1t1 – V1t2	-4.21	< .01	0.53
IEDs: V1t1 – V1t2	-12.25	< .001	1.42
Knives: V1t1 – V1t2	0.64	= .53	0.07
Other: V1t1 – V1t2	-8.95	< .001	1.12
Guns: V2t1 – V2t2	-6.03	< .001	0.70
IEDs: V2t1 – V2t2	-11.45	< .001	1.43
Knives: V2t1 – V2t2	-2.53	< .05	0.31
Other: V2t1 – V2t2	-6.17	< .001	0.84
Control group	***t*(111)**	***p***	***d***

Guns: V1t1 – V1t2	-0.21	= .84	0.02
IEDs: V1t1 – V1t2	-0.76	= .45	0.08
Knives: V1t1 – V1t2	-0.66	= .51	0.07
Other: V1t1 – V1t2	-1.26	= .21	0.13
Guns: V2t1 – V2t2	-0.67	= .50	0.09
IEDs: V2t1 – V2t2	0.71	= .48	0.07
Knives: V2t1 – V2t2	1.65	= .10	0.19
Other: V2t1 – V2t2	0.64	= .53	0.07

In summary, a large and significant training effect was found for the group who trained with XRT for three and six months compared to a control group who used another CBT for the same time. A significant training effect has been observed for all four threat categories (guns, knives, IEDs and other), whereas the extent of the effect varied between categories. A large transfer of the acquired knowledge about the visual appearance of trained objects (set A) to untrained but similar looking objects (set B) was found for the XRT training group but not for the control group. This means that training with XRT helped screeners to detect other prohibited items which were not part of the XRT training. Substantial effects of viewpoint could be observed, i.e. unusual views of prohibited objects were much harder to detect than canonical views.

4.3.7 Experiment 2

The main aim of Experiment 2 was to replicate the results of Experiment 1 at another European airport. In addition, another conventional CBT was used for the control group. Thus it could be investigated whether conventional CBTs differ from each other regarding training effectiveness compared to XRT.

4.3.7.1 Method

4.3.7.1.1 Participants

A total of 163 airport security screeners of another mid-size European airport participated in Experiment 2. All screeners conducted the X-Ray CAT 1.0.0 three times with an interval of three months between the measurements. The adaptive CBT group (XRT group) consisted of 84 screeners who conducted weekly recurrent CBT using X-Ray Tutor (XRT) CBS 2.0 Standard Edition between all three test measurements. The control group consisted of 79 screeners and they used another conventional CBT than the control group of Experiment 1. As in Experiment 1, according to the security organization and their Appropriate Authority, airport security screeners of both groups conducted about 20 min CBT per week. Analysis of XRT training use showed that on average, each screener trained 20.92 minutes (*SD* = 2.87) per week.

4.3.7.1.2 Materials and Procedure

Materials and procedure in Experiment 2 were the same as in Experiment 1. Again, all screeners took the X-Ray CAT at the beginning and after three and six months of CBT. The only difference was the CBT for the control group, which was another one than in Experiment 1. In order to avoid potential negative consequences, we decided not to mention the exact CBT product in this article for Experiment 2, neither. However, it can be mentioned that also this CBT is widely used at many airports worldwide. As the conventional CBT used in Experiment 1, this CBT has a much smaller threat image library than XRT, threat objects are not displayed in many different views, threat objects are not matched with different bags automatically on the fly, and there is no individually adaptive training algorithm.

4.3.8 Results and Discussion

This section is structured the same way as in Experiment 1. Figure 29 shows the detection performance d' for both groups and all three test measurements. As in Experiment 1, individual d´ scores were subjected to a repeated measures ANOVA with the within-participant factor measurement (first, second and third) and the between-participant factor group (XRT training group and control group). Again, there were large main effects of measurement $\eta^2 = .50$, $F(2, 322) = 163.52$, $p < .001$, group, $\eta^2 = .26$, $F(1, 161) = 56.34$, $p < .001$, and a significant interaction of measurement and group $\eta^2 = .33$, $F(2, 322) = 78.40$, $p < .001$. The large

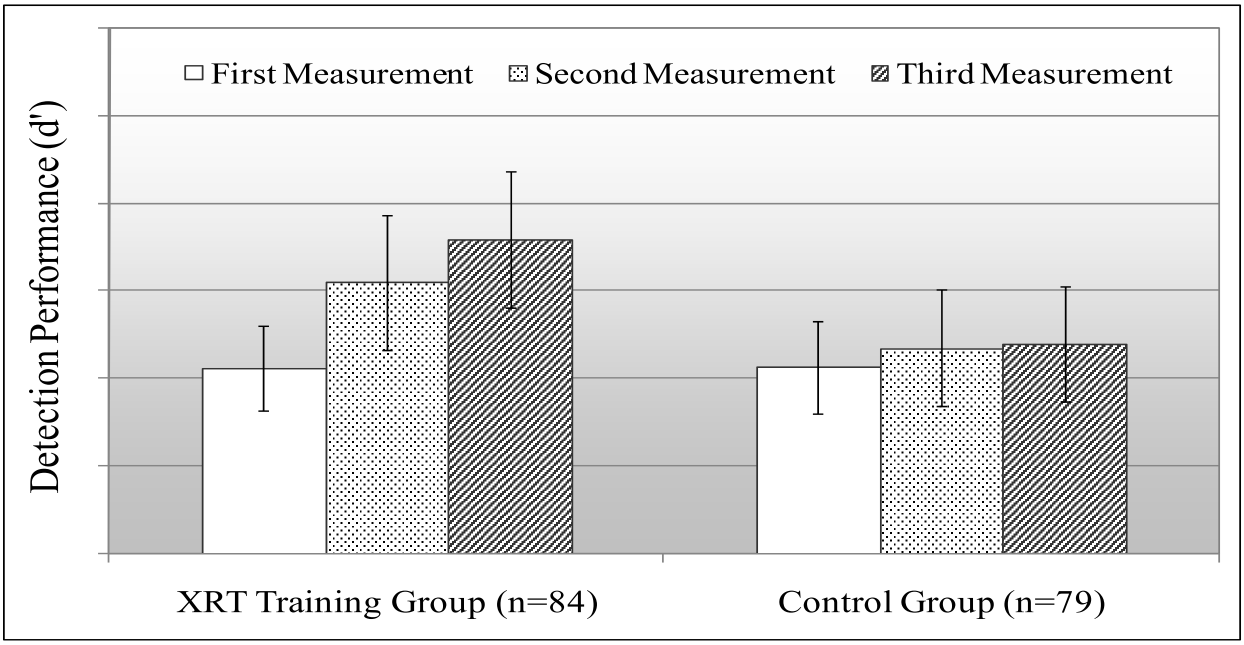

Figure 29: Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement.

interaction is consistent with Figure 29 showing a much larger performance increase as a result of training for the XRT training group when compared to the control group. This was confirmed by independent samples *t*-tests. There was no significant difference between both groups for the first measurement $t(161) = -.22$, $p = .83$, $d = 0.03$, but a highly significant difference already in the second measurement $t(161) = 6.66$, $p < .001$, $d = 1.05$ after three months of training. As in Experiment 1, additional paired-samples *t*-tests revealed significant differences for the XRT training group between all measurements. In contrast to Experiment 1, there were also significant differences for the control group between the first and second measurement, although not between the second and third measurement (see Table 13). Thus, the conventional CBT used in Experiment 2 did also result in increased detection performance although substantially less than XRT.

Table 13: Results of the *t*-tests comparing the detection performance of first (t1), second (t2) and third (t3) measurement

	***t*(83)**	***p***	***d***
XRT Training Group (t1 – t2)	-12.21	< .001	1.57
XRT Training Group (t2 – t3)	-7.07	< .001	0.65
	***t*(78)**	***p***	***d***
Control Group (t1 – t2)	-3.67	< .001	0.36
Control Group (t2 – t3)	-0.91	= .37	0.07

Figure 30 shows the detection performance of both screener groups broken up by prohibited item category and the three test measurements. Again, a clear effect of training on the detection performance can be seen for the XRT training group with the largest increase after the first three months of

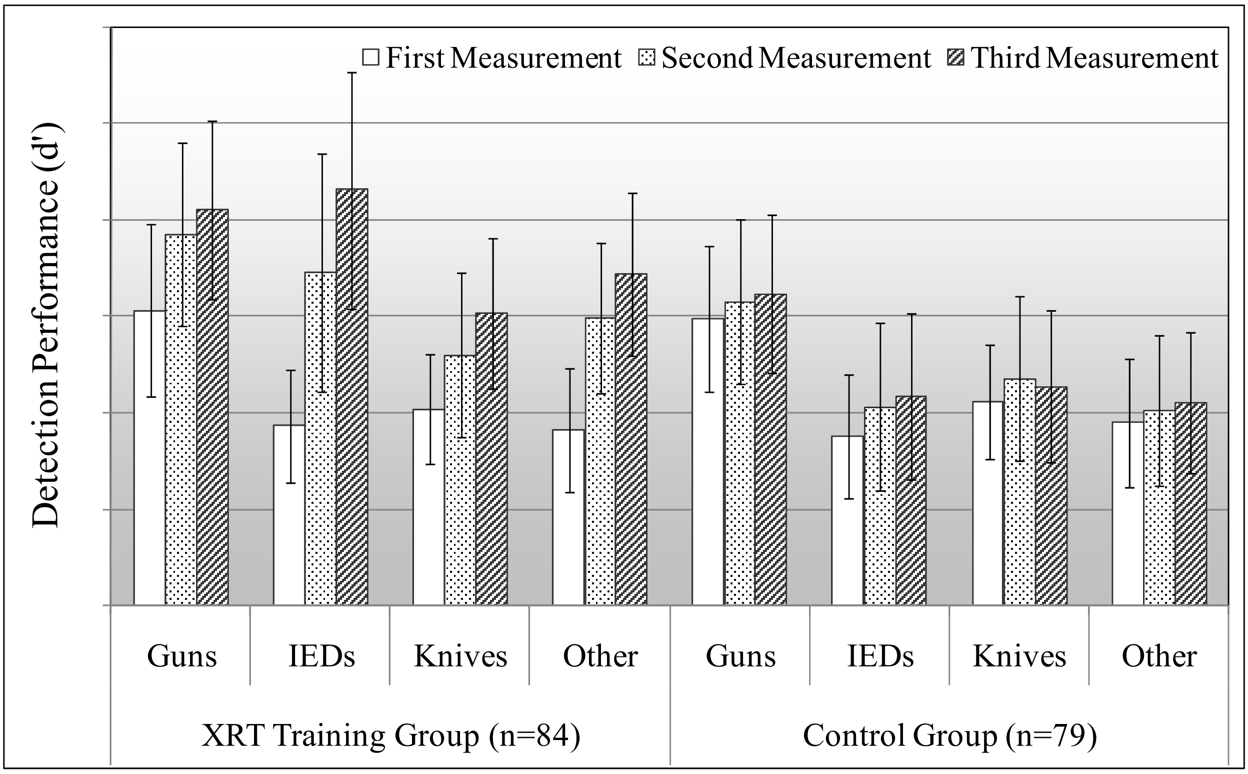

Figure 30: Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for each threat category separately

training. However, also the control group shows a slight increase in detection performance at least for the second measurement. The analysis of variance (ANOVA) with threat category as additional within-participant factor showed significant main effects and significant interactions (for details see Table 14a). The results are comparable to those in Experiment 1. Most importantly, detection of guns was best initially, while detection of IEDs was much lower. After six months of recurrent adaptive CBT, screeners of the XRT training group could detect IEDs even slightly better than guns. This nice replication of the results obtained in Experiment 1 clearly shows that IED detection is not difficult per se but only a matter of the right training. As mentioned above, all IEDs used in this study contained a detonator, wires, explosive, a triggering device and a power source. Thus our conclusions are only applicable to the detection of such multi-component IEDs. As shown in Table 15, *t*-tests between the first and second measurement revealed significant training effects for the XRT training group for all threat categories with large effect sizes (all d > .0.80). In contrast to Experiment 1, there were also significant effects for the control group, although with rather low effect sizes (all d < 0.56). Thus the conventional CBT used in Experiment 2 also resulted in performance increases although much less than XRT.

Table 14: Results of the ANOVAs in Experiment 2

	Factor	**df**	**F**	**η^2**	***p***
a)	Measurement (M)	2, 322	160.78	.50	< .001
	Threat Category (T)	3, 483	234.85	.59	< .001
	Group (G)	1, 161	64.98	.29	< .001
	M x G	2, 322	78.54	.33	< .001
	T x G	3, 483	37.63	.19	< .001
	M x T	6, 966	26.24	.14	< .001
	M x T x G	6, 966	16.67	.09	< .001
b)	Measurement (M)	2, 322	156.12	.49	< .001

	Set (S)	1, 161	58.45	.27	< .001
	Group (G)	1, 161	56.03	.26	< .001
	M x G	2,322	82.16	.34	< .001
	M x S	2, 322	8.88	.05	< .001
	S x G	1, 161	31.37	.16	< .001
	M x S x G	2, 322	15.52	.09	< .001
c)	Measurement (M)	2, 322	162.28	.50	< .001
	Set (S)	1, 161	41.88	.21	< .001
	Threat Category (T)	3, 483	231.83	.59	< .001
	Group (G)	1, 161	71.93	.31	< .001
	M x G	2, 322	84.18	.34	< .001
	M x T	6, 966	27.50	.15	< .001
	M x S	2, 322	11.42	.07	< .001
	S x G	1, 161	36.23	.18	< .001
	S x T	3, 483	33.59	.17	< .001
	T x G	3, 483	40.15	.20	< .001
	M x T x G	6, 966	16.87	.10	< .001
	M x S x G	2, 322	10.09	.06	< .001
	M x S x T	6, 966	1.48	.01	= .18
	S x T x G	3, 483	3.69	.02	< .05
	M x S x T x G	6, 966	2.64	.02	< .05

	Measurement (M)	2, 322	152.62	.49	< .001
	View (V)	1, 161	1849.85	.92	< .001
	Threat Category (T)	3, 483	216.74	.57	< .001
	Group (G)	1, 161	70.32	.30	< .001
	M x G	2, 322	80.05	.33	< .001
	M x T	6, 966	26.57	.14	< .001
	M x V	2, 322	2.99	.02	= .05
d)	V x G	1, 161	0.62	.00	= .43
	V x T	3, 483	288.98	.64	< .001
	T x G	3, 483	34.91	.18	< .001
	M x T x G	6, 966	14.95	.09	< .001
	M x V x G	2, 322	1.21	.01	= .30
	M x V x T	6, 966	2.82	.02	< .05
	V x T x G	3, 483	1.69	.01	= .17
	M x V x T x G	6, 966	1.89	.01	= .08

Table 15: Results of the *t*-tests comparing the categories between first (t1), second (t2) and third (t3) measurement

XRT training group	***t***	**df**	***p***	***d***
Guns t1 – t2	-6.01	83	< .001	0.86
IEDs t1 – t2	-12.84	83	< .001	1.74
Knives t1 – t2	-5.81	83	< .001	0.80

Other t1 – t2	-12.30	83	< .001	1.64
Guns t1 – t3	-8.19	83	< .001	1.15
IEDs t1 – t3	-20.22	83	< .001	2.70
Knives t1 – t3	-10.97	83	< .001	1.48
Other t1 – t3	-16.46	83	< .001	2.18
Control group	***t***	**df**	***p***	***d***
Guns t1 – t2	-2.19	78	< .05	0.23
IEDs t1 – t2	-3.60	78	< .01	0.42
Knives t1 – t2	-2.73	78	< .01	0.33
Other t1 – t2	-1.46	78	< .15	0.18
Guns t1 – t3	-2.72	78	< .01	0.34
IEDs t1 – t3	-4.61	78	< .001	0.56
Knives t1 – t3	-2.05	78	< .05	0.23
Other t1 – t3	-2.59	78	< .05	0.30

By an ANOVA with measurement and set as within-participant factors and group as between-participants factor, we investigated if training effects can also be shown for threat objects which were not included in the training sessions. There were main effects and interactions for all factors showing similar results as in Experiment 1 (see Table 14b for details). As in

Experiment 1, a large transfer effect was found (see Figure 31). Not only for the prohibited items of set A, which were included in the training library of XRT, but also for the untrained prohibited objects of set B, screeners of the XRT training group showed a large increase in detection performance after training. Paired-samples *t*-tests between the first and second measurement

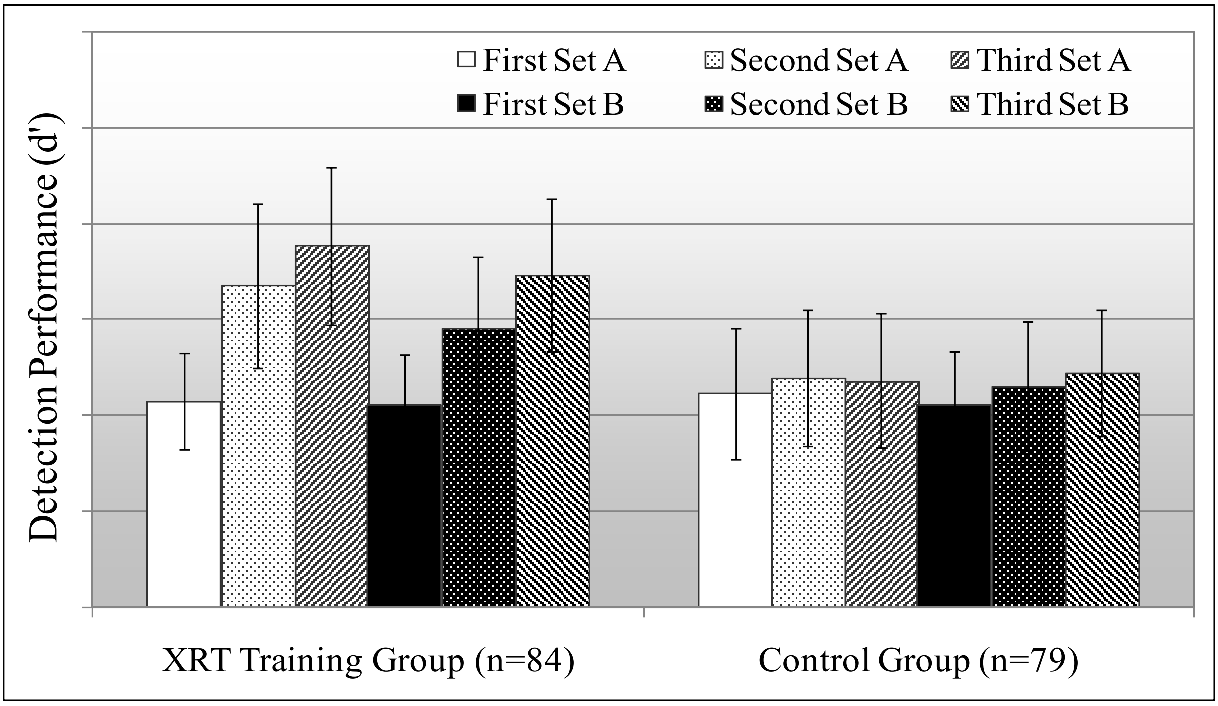

Figure 31: Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for set A and set B separately

showed training effects for both sets and also for both groups whereas again large effect sizes were found for the XRT training group and small effect sizes for the control group (trained group set A: $t(83) = -13.10$, $p < .001$, $d = 1.77$ and set B: $t(83) = -9.53$, $p < .001$, $d = 1.24$, control group set A: $t(78) = -2.32$, $p < .05$, $d = 0.24$ and set B: $t(78) = -3.00$, $p < .01$, $d = 0.32$). Pairwise *t*-tests showed no significant difference in the difficulty of set A and Set B for both groups at the first measurement (XRT training group: $t(83) = 1.16$, $p = .25$, $d = 0.10$, control group: $t(78) = 1.93$, $p = .06$, $d = 0.19$).

Figure 32 includes also the threat category in the analysis. Paired samples

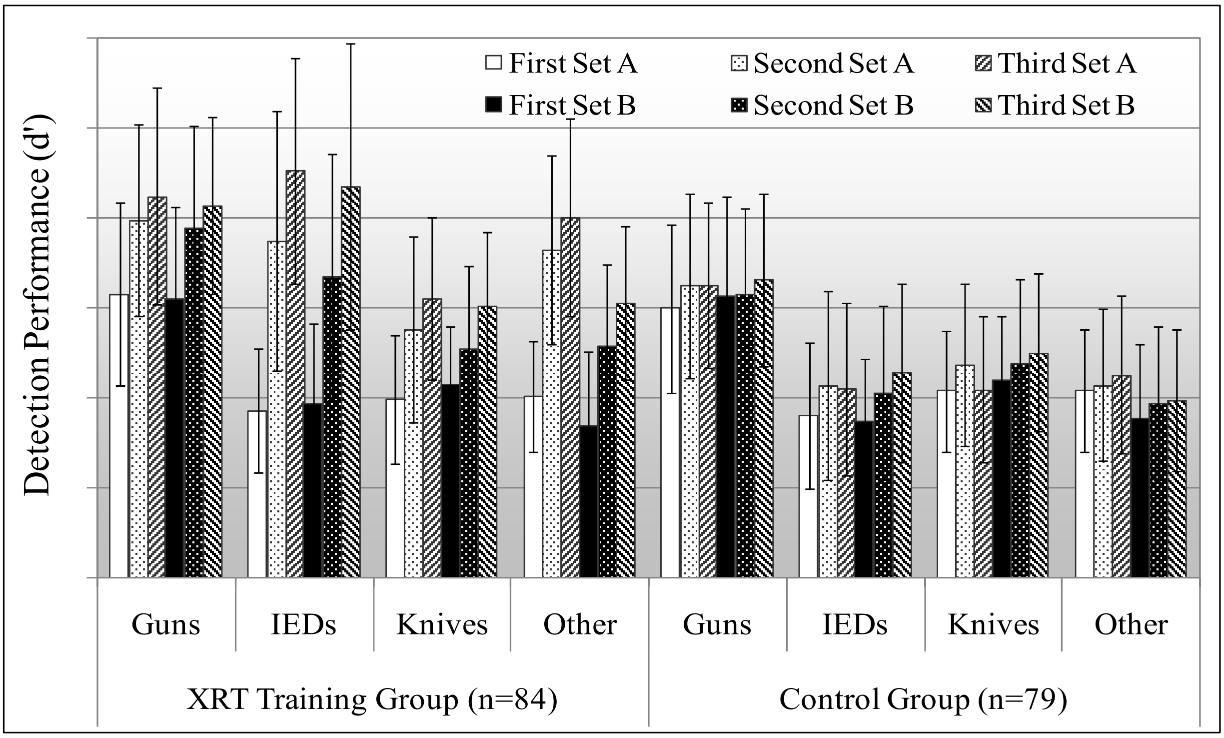

Figure 32: Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for set A and set B and each

t-tests were calculated in order to investigate if the training effect between the first and second measurement was significant for each category in both sets for the XRT training group. Results revealed significant effects for all categories in each set ($p < .01$, $d = 0.51$ for knives in Set B, $p < .001$, $d > 0.74$ for all other categories). Thus, as in Experiment 1, XRT resulted in large detection performance increases even for prohibited objects that are not part of the XRT image library (X-Ray CAT image set B). For the control group the difference between the first and *third* measurement was calculated in order to maximize the chances for finding a significant training effect. The following *t*-tests were significant: IEDs for both sets, knives only for set A, and other threat objects for both sets ($p <.05$, $d > 0.23$). All other values were not significant ($p > .06$, $d < 0.28$) and reveal no effect of training between the different measurements.

As in Experiment 1, individual d´ scores were subjected to an extended ANOVA with the within-participant factors measurement, X-Ray CAT image set, threat category and the between-participants factor group. All main effects and interactions were significant except the interaction between measurement, set and threat category (see Table 14c for details). In contrast to Experiment 1 the ANOVA revealed a main effect of set and significant interactions with set. However, as can be seen in Figure 31 they were rather small, which implies large transfer effects. As in Experiment 1 the results clearly show a training effect for each category and in both sets. This is consistent with the results of the *t*-tests explained above. The training effect that was found for the control group revealed itself also in the sets, that is, there was a transfer effect for the control group, too.

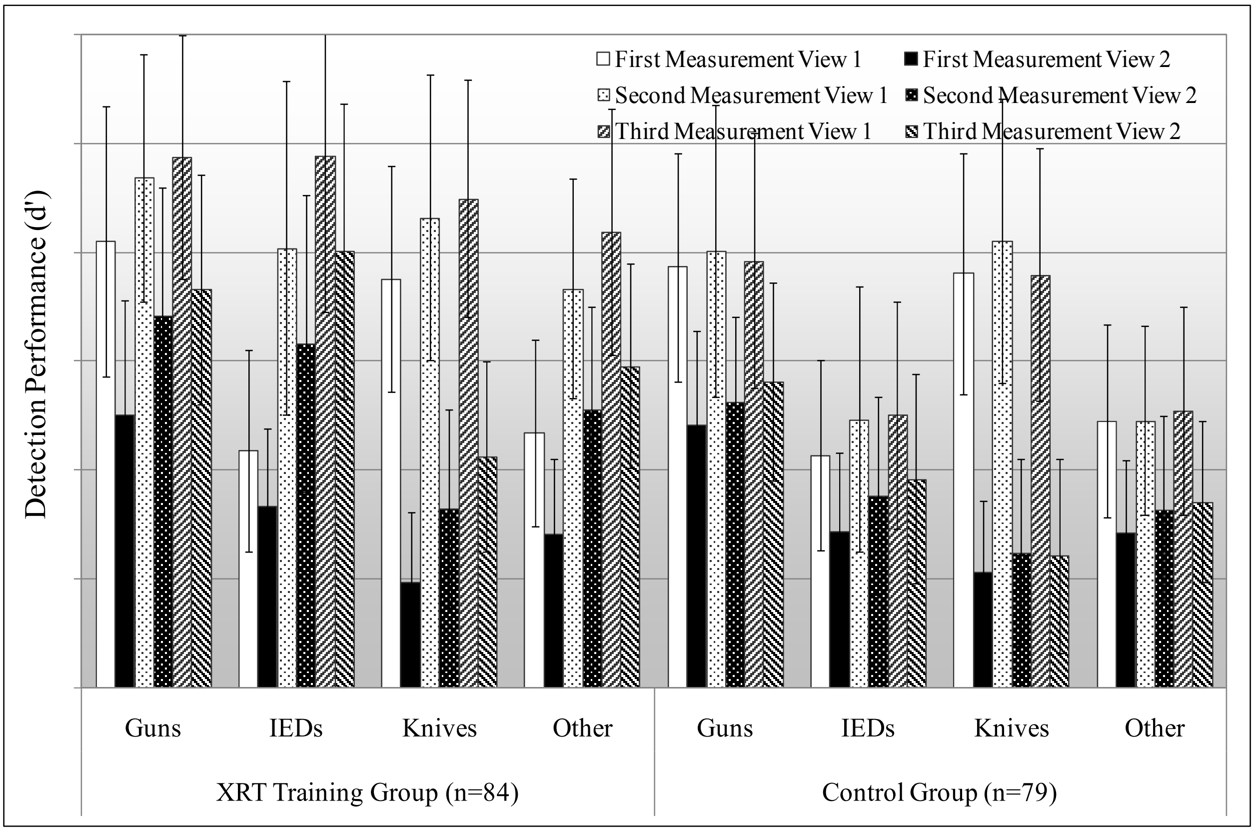

Figure 33: Detection performance with standard deviations for the XRT training group vs. the control group comparing first, second and third measurement for both views and each threat category separately

Last, the effect of viewpoint was investigated calculating a four-way ANOVA. Results show clear main effects of measurement, view, threat category and group. For details on interactions please refer to Table 14d. Detection performance is clearly much higher for objects that are shown in the easy view (View 1) than for the objects that are shown from an unusual viewpoint (see Figure 33). This effect is valid for all threat categories and for the XRT training group as well as for the control group. However, the viewpoint effect is not the same for different threat categories. The graphs in Figure 33 suggest that the largest viewpoint effect can be observed for the detection of knives, the smallest one for IEDs.

As in Experiment 1, pairwise *t*-tests showed a significant increase in detection performance at the second measurement for both views for the XRT training group for all four threat categories ($p < .01$, $d > 0.49$. For the easy view, the control group showed a significant effect for IEDs only ($p < .05$, $d = 0.32$), all other *t*-tests were not significant ($p > .07$, $d < 0.25$). For the difficult view all *t*-test with one exception were significant for the control group ($p < .05$, $d > 0.26$). Only the training effect of knives in the rotated view was not significant $p = .07$, $d = 0.24$ (see Table 16 for details). But the results show that although some significant effects in the control group were observed, effect sizes were small compared to those of the XRT training group.

Table 16: Results of the *t*-tests comparing the detection performance of the four categories for easy view (V1) and difficult view (V2) between the first (t1) and second (t2) measurement

XRT training group	***t*(83)**	***p***	***d***
Guns: V1t1 – V1t2	-3.59	< .01	0.49

IEDs: V1t1 – V1t2	-10.93	< .001	1.51
Knives: V1t1 – V1t2	-4.35	< .001	0.48
Other: V1t1 – V1t2	-9.79	< .001	1.42
Guns: V2t1 – V2t2	-5.46	< .001	0.82
IEDs: V2t1 – V2t2	-9.99	< .001	1.45
Knives: V2t1 – V2t2	-5.79	< .001	0.88
Other: V2t1 – V2t2	-10.33	< .001	1.40
Control group	***t*(78)**	***p***	***d***
Guns: V1t1 – V1t2	-1.07	= .29	0.13
IEDs: V1t1 – V1t2	-2.64	< .05	0.32
Knives: V1t1 – V1t2	-1.87	= .07	0.25
Other: V1t1 – V1t2	-0.05	=.96	0.01
Guns: V2t1 – V2t2	-2.35	< .05	0.26
IEDs: V2t1 – V2t2	-3.24	< .01	0.41
Knives: V2t1 – V2t2	-1.81	= .07	0.24
Other: V2t1 – V2t2	-2.11	< .05	0.28

In summary, very similar results as in Experiment 1 have been found in Experiment 2. A large and significant training effect was observed for the group who trained with XRT compared to a control group who used a conventional CBT for the same time. A significant training effect has been observed for all four categories (guns, knives, IEDs and other) for the XRT training group, whereas the effect size varied between categories. Also a large transfer of the acquired knowledge about the visual appearance of trained objects (set A) to untrained but similar looking objects (set B) was found for the XRT training group. Additionally a viewpoint effect could be observed which shows that unusual views of forbidden objects are much harder to detect than canonical views. In contrast to Experiment 1, the control group also showed increases of detection performance, which implies that the conventional CBT used in Experiment 2 is more effective than the one used in Experiment 1 (although still much less effective than XRT). Moreover, there was also a transfer effect for the control group.

4.3.9 General Discussion

The first aim of this study was to investigate how well airport security screeners can detect guns, knives, IEDs and other prohibited items in x-ray images of passenger bags. Two experiments conducted at two European airports provided very similar results. A computer-based test (X-Ray CAT) was conducted before and after three and six months of weekly (about 20 min per screener) CBT at each airport. The first measurement revealed that guns were detected best, followed by knives, other prohibited items and IEDs. In both experiments and airports, one group used an adaptive CBT (X-Ray Tutor, XRT) with individually adaptive algorithms, a large library of prohibited items depicted in a variety of different views, and automatically created prohibited item to bag combinations (see Schwaninger, 2004 for details). The other group used a conventional CBT system with no adaptive algorithms, a smaller image library, and fixed combinations of threat items in bags. While XRT was used in both experiments and airports, two different conventional CBT systems were used for the control groups of Experiment 1 (airport 1) and Experiment 2 (airport 2). At both airports, XRT training group results revealed a training effect for all types of threat objects (guns, knives, IEDs, and other prohibited items). However, effect sizes differed remarkably for the four categories. While guns were detected best and IEDs were detected worst at the beginning, IED detection of the XRT training group was as good as or even slightly better than gun detection after several months of training. This shows that the detection of IEDs is not difficult per se, but rather depending on the training of screeners. Note that all IEDs used in this study contained a detonator, wires, explosive, a triggering device and a power source. Therefore, these conclusions are only applicable to the detection of such multi-component IEDs. However, a large training effect for IEDs can be expected because they are usually not encountered at airport security checkpoints and therefore not known to screeners without enhanced training in IED detection. The relatively large training effect for the category “other” which includes self defense gas spray, electric shock devices etc. might also be explained by less on the job exposure of these prohibited items. In a study with hold baggage screeners, large training effects for IEDs were also found, which is very consistent with results of this study (Schwaninger and Hofer 2004). In contrast to IEDs and other prohibited items, guns seem to be well known by screeners either because of their typical shape or the frequency by which they are encountered at the airport security screening checkpoint (e.g., toy guns). Therefore, detection performance before training is already high for guns and a large improvement is impossible. It is also noticeable that detection for knives showed the smallest training effect in both experiments. Although the detection was at the baseline measurement higher than for IEDs and other prohibited items, after six months of training screeners’ performance was poorest for knives. On average, knives are smaller than IEDs and other threat items and show less diagnostic features. This might be a reason for the lower detection performance increase for this threat category.

While training with XRT resulted in large training effects, the tested conventional CBT systems were less effective. In Experiment 1, there were no training effects at all, while only small training effects were observed for the conventional CBT system used in Experiment 2. This could be due to one or a combination of the following reasons: First, the conventional CBT systems tested in this study do not feature individually adaptive training algorithms like XRT (see Schwaninger, 2004 for details). Second, in contrast to XRT, the conventional CBT systems did not contain such a large image library with many prohibited items depicted from a variety of different viewpoints. Third, while in XRT prohibited items are blended into x-ray images of passenger bags on the fly using scientifically validated and individually adaptive algorithms based on image measurement as described in Schwaninger et al. (2007), the conventional CBT systems used in Experiment 1 and 2 have only fixed combinations of prohibited items in bags. Finally, we had to rely on the statement of the appropriate authority and the security companies regarding the amount of training that was conducted by screeners of the control group and the XRT training group, which should have been on average 20 min per week per screener. Analysis of XRT training data showed, that this was clearly fulfilled for screeners of the XRT training group at both airports.

Since the X-Ray CAT is composed of two comparable (similar looking) sets (set A and set B) whereof only the threat objects of set A were included into the XRT training system, transfer effects can be tested, i.e. whether training with certain prohibited items helps increasing detection of other prohibited items that are not contained in the training. Overall, the comparison of the two sets A and B at the baseline measurement (before training) shows no significant difference. However, in Experiment 1 there was a slight difference for the control group between the two sets indicating that the two sets are not exactly equal in terms of image difficulty for this sample. But this possible objection to the transfer effect can be disapproved with two arguments: first, the effect size was only small according to the conventions by Cohen (1988) and second, only one of the two control groups showed a significant difference. Therefore, the transfer effect in the results of the XRT training group can be attributed to the training of set A only. The small training effect for the control group in Experiment 2 is also reflected in the detection increase of both sets after training. Although the conventional CBT system of this control group did not contain any objects from the test. The training with this training system apparently also led to a transfer of the knowledge to the objects in the test. In another study it would be interesting to compare the objects that are comprised in the two training systems used by the control groups regarding their similarity to the test objects. Contrary to our results, Smith, Redford, Gent, and Washburn (2005) found a large decrease in screeners' detection performance when specific trained objects were replaced with new images belonging to the same categories (see also Smith, Redford, Washburn, and Taglialatela 2005). According to these authors, improvement in screening performance is attributable only to specific-token familiarity that developed for the original images and not to a category generalization. They state constraints on categorization and the use of category-general information when humans face visual complexity and have to identify targets within it. Our results can be interpreted in support of generalization of visual learning in x-ray image interpretation.

However, it might be possible that the objects of the untrained set in our study are so similar to the trained objects that a specific-token familiarity led to the detection performance increase and not a true generalization effect. The lacking transfer effect in knives would along these lines mean that the objects in set A and set B are not similar enough in shape to generate a specific-token familiarity. Therefore only the learnt objects could generate a training effect but not the unlearnt ones. For Schwaninger and Hofer's (2004) findings of a large increase in detection performance of IEDs after recurrent CBT with other members of the category than those included in the test, it would mean, that those objects were very similar in order to create a specific-token familiarity and therefore a training effect.

In both Experiments a large viewpoint effect was also revealed. This is consistent with view-based theories of object recognition (for reviews see for example Tarr and Bülthoff, 1995, 1998; Graf et al., 2002; Hayward, 2003). After training, easy and difficult views were recognized much better. Interestingly, there was no significant interaction between measurement and viewpoint, i.e. although training resulted in improved performance for difficult views, the viewpoint effect (impairment for unusual vs. canonical views) remained stable even after six months of training. However, it must be pointed out that the XRT training algorithm only provides the screeners with unusual views of objects once a screener can detect a prohibited item well when depicted from easy perspective. That is, when screeners start to train with XRT all threat objects are shown in easy views. Only if these objects are detected reliably, the difficulty level is increased for a certain threat item by showing it in more difficult views (Schwaninger, 2004). Thus, it is unclear whether a significant interaction between viewpoint and measurement would have been observed if the training duration would have been increased (e.g., to one year). The conclusion stands to reason that recognition of forbidden objects in x-ray images is dependent on exposure which has very important implications for an adaptive training system. It has been assumed that different views of each object become associated with one another during object rotation, either through active learning or through passive experiencing of the successive appearance of nearby views (Földiák, 1991; Stryker, 1991). Hence, it is important that during training screeners are getting feedback which forbidden object has been detected or missed. This feedback shows the photograph and also the x-ray image of that forbidden object always in the canonical view whereas the forbidden object merged into a bag is presented in different viewpoints. This leads to an association between an unusual view of an object and the canonical view which results in a sequential pairing of these views with each other (Wang, Obama, Yamashita, Sugihara and Tanaka, 2005). This association, which forms during learning, is thought to underlie object recognition ability across changes in viewing angle (Palmeri and Gauthier, 2004).

For our future studies, it could also be interesting to increase the interval between the end of training and the testing of training transfer, as corresponding literature usually tests transfer of training after a considerable period of time in order to measure the stability of the transfer (e.g., Saks and Belcourt, 2006). In any case, our findings show that the knowledge about the visual appearance of forbidden objects, which airport security screeners acquire during recurrent CBT, can be transferred to similar looking, but not previously seen objects and also the effect that rotated views are much harder to detect can be decrease with training. To make sure that objects are well detected it is important that a large and representative image library of prohibited objects is used and that these objects are learned from different viewpoints. Additionally the library should be updated constantly to adapt to new threats. Overall, this study has shown that adaptive CBT can be a powerful tool to increase screeners' x-ray image interpretation competency in an efficient and effective way.

4.3.10 Acknowledgments

This research was financially supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403). Many thanks to Zurich State Police, Airport Division, for their help in creating the stimuli and the good collaboration for conducting parts of the study.

4.3.11 References

Bülthoff, H.H., Edelman, S.Y., Tarr, M.J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex,* 3:247–260.

Biederman, I. (1987). Recognition-By-Components: a theory of humanimage understanding. *Psychological Review,* 94:115-147.

Cohen, J. (1988). Statistical *power analysis for the behavioural sciences*. Erlbaum, Hillsdale, New York.

Földiák, P. (1991). Learning invariance from transformation sequences, *Neural Computation,* 3:194-200.

Graf, M., Schwaninger, A., Wallraven, C., Bülthoff, H.H. (2002). Psychophysical results from experiments on recognition and categorisation. Information Society Technologies (IST) programme, *Cognitive Vision Systems – CogVis,* (IST-2000-29375).

Green, D.M., Swets, J.A. (1966). *Signal Detection Theory and Psychophysics,* Wiley, New York.

Hardmeier, D., Hofer, F., Schwaninger, A. (2006). The role of recurrent CBT for increasing aviation security screeners' visual knowledge and abilities needed in x-ray screening. *The 4^{th} International Aviation Security Technology Symposium*, Washington, D.C., USA, November 27 – December 1, 2006.

Hayward, W.G. (2003). After the viewpoint debate: where next in object recognition? *Trends in Cognitive Sciences*, 7 (10):425 – 427.

Hummel, J.E., Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99(3), 480-517.

Koller, S., Schwaninger, A. (2006). Assessing X-ray image interpretation competency of airport security screeners. *Proceedings of the 2nd International Conference on Research in Air Transportation*, ICRAT 2006 Belgrade, Serbia and Montenegro, June 24-28, 2006:399-402.

Kosslyn, S.M. (1994). Image and brain. *The resolution of the imagery debate*. MIT Press, Cambridge, Massachusetts.

Lawson, R. (1999). Achieving visual object constancy across plane rotation and depth rotation. *Acta Psychologica,* 102:221–245.

Logothetis, N.K., Sheinberg, D.L. (1996). Visual object recognition. *Annual Review of Neuroscience,* 19:577–621.

Lowe, D.G. (1985). *Perceptual organization and visual recognition*. Boston: Kluwer Academic Publishing.

Lowe, D.G. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355-395.

Marr, D., Nishihara, H.K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. Proc. R. Soc. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 200:269–294.

McCarley, J.S., Kramer, A.F., Wickens, C.D., Vidoni, E.D., Boot, W.R. (2004). Visual skills in airport screening. *Psychological Science,* 15 (5):302-306.

Palmer, S.E., Rosch, E., Chase, P. (1981). Canonical perspective and the perception of objects. In: Long I, Baddeley A (eds) *Attention and Performance IX*. Erlbaum, Hillsdale, N.J.

Palmeri, T.J., Gauthier, I. (2004). Visual object understanding. Nature Review Neuroscience, 5:291–303.

Saks, A.M., Belcourt, M. (2006). An Investigation of Training Activities and Transfer of Training in *Organizations. Human Resource Management,* 45 (4)*:*629-648.

Peissi, J., Tarr, M.J. (2007). Visual Object Recognition: Do We Know More Now Than We Did 20 Years Ago? *Annual Review of Psychology* 58: 75 – 96.

Poggio, T., Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature* 343(6255):263-266.

Schwaninger, A. (2003). Training of airport security screeners. *AIRPORT,* 05/2003:11-13.

Schwaninger, A. (2004). Computer based training: a powerful tool to the enhancement of human factors. *Aviation Security International,* FEB/2004:31-36.

Schwaninger, A. (2005). Object recognition and signal detection. In: Kersten B (ed) *Praxisfelder der Wahrnehmungspsychologie*. Huber, Bern.

Schwaninger, A. (2005b). X-ray Imagery: enhancing the value of the pixels. *Aviation Security International.* Oct 2005:*16*-21.

Schwaninger, A., Hardmeier, D., Hofer, F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. *IEEE Aerospace and Electronic Systems,* 20(6):29-35.

Schwaninger, A., Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in x-ray screening. In: Morgan K, Spector MJ (eds) The Internet Society 2004, *Advances in Learning, Commerce and Security*. WIT Press, Wessex.

Smith, J.D., Redford, J.S., Gent, L.C., Washburn, D.A. (2005). Visual search and the collapse of categorization. *Journal of Experimental Psychology*: General 134 (4):443-460.

Schwaninger, A., Michel, S., Bolfing, A. (2007). A statistical approach for image difficulty estimation in x-ray screening using image measurements*. Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization,* ACM Press, New York, USA*:*123-130.

Schwaninger, A., Wetter, O.E., Hofer, F. (2007). Adaptive computer-based training increases on the job performance of x-ray screeners. *Proceedings of the 41st Carnahan Conference on Security Technology*, Ottawa, October 8-11, 2007.

Smith, J.D., Redford, J.S., Washburn, D.A., Taglialatela, L.A. (2005). Specific-token effects in screening tasks: possible implications for aviation security. *Journal of Experimental Psychology*: Learning, Memory, and Cognition 31 (6):1171-1185.

Stryker, M.P. (1991). Temporal associations. *Nature* 354:108–109.

Tarr, M.J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review,* 2:55–82.

Tarr, M.J., Bülthoff, H.H. (1995). Is human object recognition better described by geon-structural-descriptions or by multipleviews? *Journal of Experimental Psychology: Human Perception and Performance*, 21(6):1494–1505.

Tarr, M.J., Bülthoff, H.H. (1998). Image-based object recognition in man, monkey and machine. In: Tarr MJ, Bülthoff HH (eds) *Object recognition in man, monkey, and machine*. MIT Press, Cambridge, MA.

Tarr, M.J., Pinker, S. (1989) Mental rotation and orientation-dependence in shape-recognition. *Cognitive Psychology*, 21(2): 233-282.

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. In: Tarr M, Bülthoff H (eds) *Object Recognition in Man, Monkey and Machine*, MIT Press, England.

Verfaillie, K. (1992). Variant points of view on viewpoint invariance. *Canadian Journal of Psychology,* 46:215–235.

Wang, G., Obama, S., Yamashita, W., Sugihara, T., Tanaka, K. (2005). Prior experience of rotation is not required for recognizing objects seen from different angles. *Nature Neuroscience,* 8 (12):1768–1775.

5 Benefits and Costs of New X-Ray Image Technology

5.1 Do multi-view X-ray systems improve X-ray image interpretation in airport security screening?

5.1.1 Abstract

X-ray screening of passenger bags is one of the core elements in aviation security in order to prevent terrorist attacks. Large investments have been made into new technologies, for example in multi-view X-ray systems. Because of several X-rays, multi-view systems provide more than one X-ray image of the same passenger bag and hence present the security screener multiple perspectives of that bag. In this study, we evaluated the benefit of multi-view X-ray systems compared with state-of-the-art single-view X-ray systems. Single- and multi-view X-ray images of passenger bags were presented to 32 novices who had to decide if the bag contained a prohibited item or not. The results show that multi-view X-ray systems lead to a higher detection performance of prohibited items in difficult conditions, such as when it is rotated in a non-canonical manner or superimposed by other objects. Additionally, the results indicate an increase of the reaction time for performing the screener's task with multi-view in comparison with single-view X-ray systems. A specific training for airport's security screeners might increase the advantages and reduce the disadvantages of multi-view X-ray systems.

5.1.2 Introduction

The increasing threat of terrorist attacks in recent years has led to large investments into technological enhancements in aviation security. One main focus continues to be the improvement of the process of X-ray screening of passenger bags to prevent forbidden objects getting past the security checkpoint. Some airports have started using technologies of the newest generation for the process of cabin baggage screening, such as multi-view X-ray systems. Certain multi-view X-ray systems even provide automated detection of explosives, leading to a substantial improvement of security. However, for the detection of other types of prohibited items, airports still rely on human operators (airport security screeners) who visually inspect X-ray images.

The task of threat object detection depends on knowledge-based and image-based factors (Hardmeier et al., 2005; Schwaninger et al., 2004). Knowledge-based factors refer to knowing which items are prohibited and what they look like in X-ray images of passenger bags. Some objects look quite different in X-ray images than in reality, for example an electric shock device, which is difficult to differentiate from ordinary objects (see Figure 34). Others, such as Improvised Explosive Devices (IED), are rarely seen in everyday life as well as at the security checkpoint and are, therefore, difficult to recognize without the appropriate training.

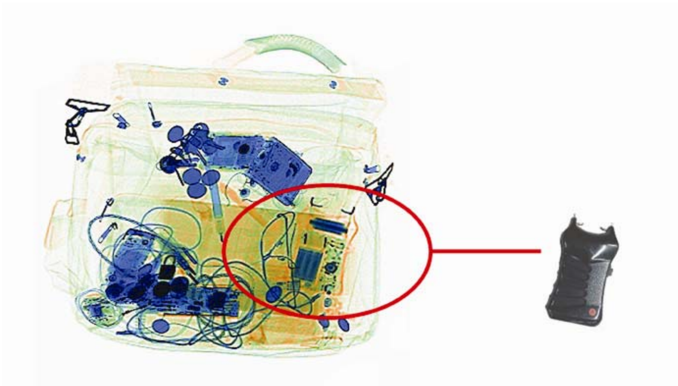

Figure 34: Illustration of the impact of knowledge-based factors. Some forbidden objects such as the electric shock device shown above look quite different in X-ray images than in reality.

Furthermore, image-based factors play an important role in threat object detection. These can be attributed to the visual abilities of a person, that is, how he or she copes with image difficulty. Schwaninger (2003) described three image-based factors: rotation, superposition, and bag complexity (Figure 35).

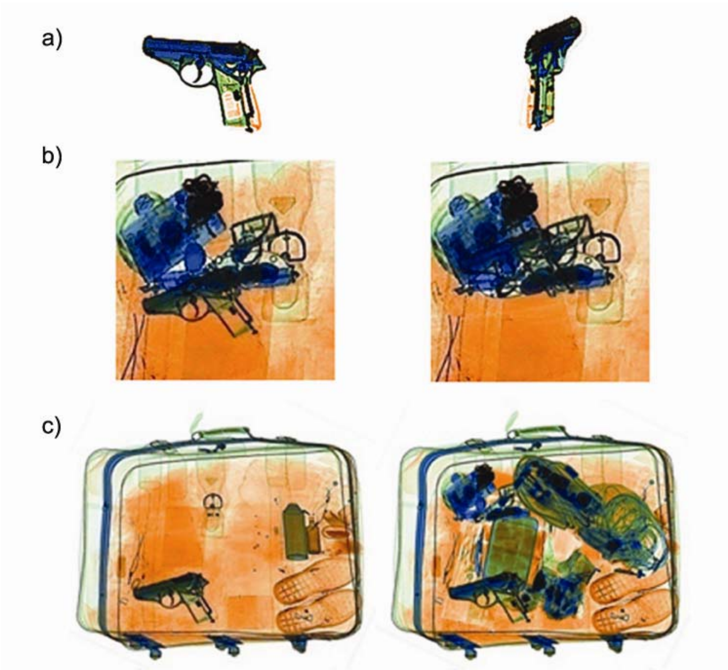

Figure 35: Image-based factors with an impact on threat detection performance: (a) easy and difficult rotation, (b) low and high superposition, and (c) low and high bag complexity.

The fact that object recognition often yields strong effects of viewpoint caused by the rotation of an object (e.g. Bülthoff & Edelman, 1992; Graf et al., 2002; Tarr & Bülthoff, 1995, Tarr & Bülthoff, 1998), is essential for X-ray image interpretation. In general, X-ray images of forbidden objects are difficult to recognize when depicted from an unusual viewpoint and when diagnostic features are not visible.

Another important factor contributing to image difficulty is the superposition of the threat object by other objects in a bag. The effect of superposition refers to the impairment of figure-ground segregation. If a threat object, such as a knife, is superimposed by high density material, it becomes more difficult to recognize the characteristic shape of the object.

Furthermore, the complexity of a bag, determined by the number and type of objects in the bag, has a significant influence on the detection performance. Recently, the effects of the image-based factors rotation, superposition and bag complexity have been replicated, and statistical algorithms for estimating image difficulty have been developed (Schwaninger et al. 2007b).

Besides coping with knowledge-based and image-based factors, the screener has to detect a threat object in a limited amount of time. During rush hour at the security checkpoint, screeners often have only a few seconds to visually inspect the X-ray images of passenger bags.

There is evidence that perceptual training can help to improve the ability to segment objects from cluttered visual scenes (e.g. Brady & Kersten, 2003; Kourtzi et al., 2005; Kovacs et al., 1999; Li & Gilbert, 2002; Yi et al., 2006). Furthermore, it has been shown that a specific X-ray screening training increases the X-ray image interpretation competency and decreases reaction times (Koller et al., 2007, Koller et al., 2008; McCarley et al., 2004; Michel et al., 2007; Schwaninger et al., 2007a). Such a computer-based object recognition training affects mainly knowledge-based factors and the detection of rotated objects (Schwaninger et al., in press). Through training, airport's security screeners learn which objects are prohibited and how they appear in X-ray images. The screeners also store different and often unfamiliar views of the objects in visual memory (Michel et al., 2007).

Effects of image-based factors might also be diminished by recent technological improvements. Multi-view X-ray systems produce two or more images of one object due to several X-rays. In airport security screening this means that the decision of the screener, if a passenger bag is OK or not, is supported by multiple images of the same bag. Threat objects that are rotated in a non-canonical manner or superimposed by other objects in the bag might be recognized easier when a second, e.g. 90 degree rotated view, is available. Assuming that visual search needs more time for multiple images than for a single image, reaction times are probably longer for multi-view than for single-view X-ray systems.

In this study, we conducted an experiment in order to investigate what impact multiple views have on the detection performance as well as on the reaction time of a screening person. In order to eliminate the influence of knowledge-based factors, we conducted the experiment with novices who had no prior expertise in X-ray screening, and we presented only threat objects, whose appearance is relatively common in everyday life, such as guns and knives. We will examine the impact of the X-ray system on the introduced image-based factors rotation and superposition, as well as on the reaction time.

Our hypothesis is that with multi-view X-ray systems, the detection performance increases especially for difficult conditions, i.e. a difficult rotation or a high superposition of the threat item, based on the assumption that a second view might eliminate or at least reduce the negative influence of the introduced image-based factors. Additionally, reaction times probably increase for multi-view X-ray systems in comparison with single-view X-ray systems because visual search needs more time for multiple images, as mentioned above.

5.1.3 Method

5.1.3.1 Participants

Thirty-two undergraduate students from the University of Zurich volunteered to participate in this study; 20 females and 12 males with age ranging from 19 to 50 years (*M* = 26.69 years, *SD* = 6.35 years). All participants reported normal or corrected-to-normal vision. They were all naive with regard to the hypotheses under investigation.

5.1.3.2 Material

For our experiment we constructed a test consisting of X-ray images of passenger bags by using a multi-view X-ray system. The approximate directions of the two X-ray beams are illustrated schematically in Figure 36.

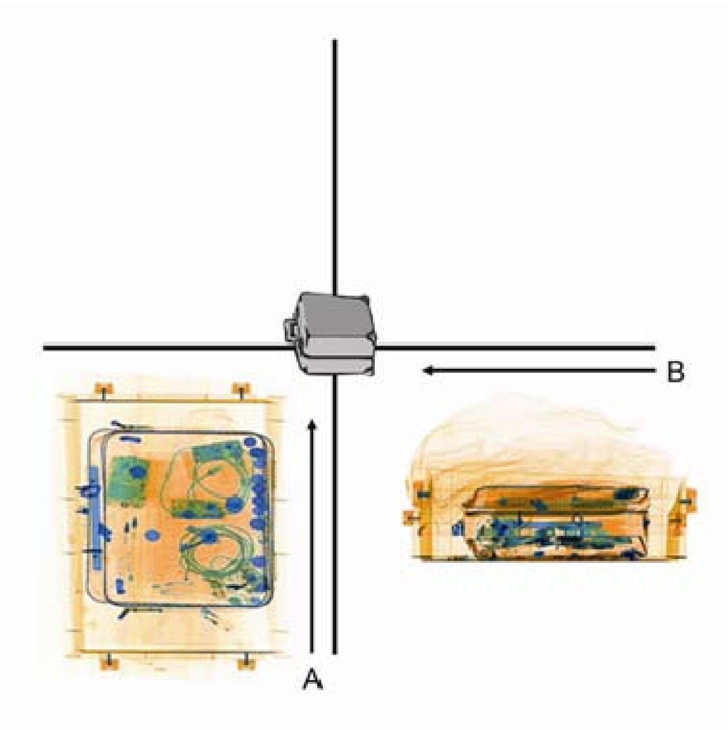

Figure 36: Schematic illustration of the approximate directions of the two X-ray beams A and B.

One half of the total of 128 bags in our experiment had a low, and the other half a high bag complexity. We calculated the level of bag complexity using the formula for transparency, which reflects the extent to which X-rays are able to penetrate objects in a bag, as described by Schwaninger et al. (2007b):

$$TR = \frac{\sum_{x,y}(I_N(x,y) < threshold)}{\sum_{x,y}(I_N(x,y) < 255)}$$

$I_N(x, y)$ denotes the pixel intensities, whereas threshold is the pixel intensity threshold beneath which the pixels are counted.

All of the bags were used twice, once combined with a prohibited item, and once without any threat object. The threat objects have been captured by experts of Zurich State Police, Airport division. Because we tested novices who were not trained in recognizing unfamiliar objects like IED, we only used eight guns and eight knives to eliminate effects of knowledge-based factors. Each threat item was presented in two different rotations. The easy rotation shows the object from a canonical perspective (Palmer et al., 1981) as judged by two security experts who captured the stimuli. The difficult rotation shows the threat item 85 degree rotated horizontally or vertically relative to the canonical view.

Every threat item was combined with a bag in a manner that the degree of superposition by other objects was low, and with another bag in a manner that the degree of superposition by other objects was high. For this purpose we used the formula for superposition as described by Schwaninger et al. (2007b):

$$SP = \sqrt{\sum_{x,y}(I_{SN}(x,y) - I_N(x,y))^2}$$

The function computes the difference between the pixel intensity values of the bag image with the threat object ($I_{SN}(x, y)$) minus the pixel intensities of the corresponding harmless bag ($I_N(x, y)$).

Every image was once presented as single-view and once as multi-view trial. The single-view condition consisted of one view (see Figure 37, X-ray A), i.e. one image was shown, the multi-view condition consisted of two views (see Figure 37, X-rays A and B), i.e. two images were presented at once on the same screen. The size of the bag images was consistent regardless of the condition. As shown in Figure 37b, the first view (left side) of the multi-view trial was identical with the corresponding single-view trial, whereas the second view (right side) was an image of the same bag from another angle.

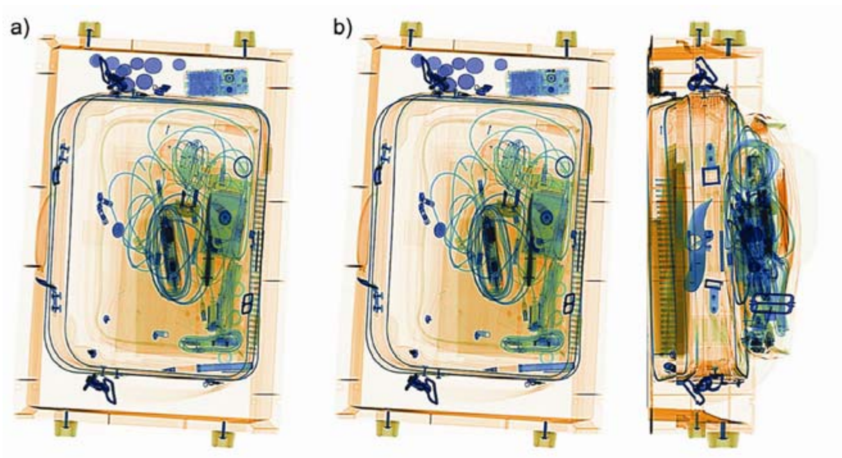

Figure 37: (a) A single-view trial containing a knife in a difficult rotation and high superposition and (b) the corresponding multi-view trial. In the second view of the multi-view trial, the rotation of the threat object is easy and the superposition low.

The variation of the image-based factors rotation, superposition, and bag complexity resulted in eight variations for the single-view condition, and, therefore, also for the first view of the multi-view condition. In the second view, we varied the factors rotation and superposition again, resulting in another four variations for the second view: easy rotation and low superposition, easy rotation and high superposition, difficult rotation and low superposition, and difficult rotation combined with high superposition.

These four variations of the second view have been balanced throughout four groups of participants (cp. Table 17). Overall, 512 trials were presented: 16 (threat objects) * 2^3 (image-based factors rotation, superposition, bag complexity) * 2 (single-/multi-view X-ray system) * 2 (bag with/without threat item).

Table 17: Design of multi-view and single view trials

Multi-view variations of bags containing a threat item					
		Second view			
Bag and its complexity	First view	g1	g2	g3	g4
low					
1	t1r1s1	t1r1s1	t1r2s1	t1r1s2	t1r2s2
2	t2r1s1	t2r2s1	t2r1s2	t2r2s2	t2r1s1
3	t3r1s1	t3r1s2	t3r2s2	t3r1s1	t3r2s1
4	t4r1s1	t4r2s2	t4r1s1	t4r2s1	t4r1s2
…	…	…	…	…	…
high					
…	…	…	…	…	…
128	t16r2s2	t16r2s2	t16r1s1	t16r2s1	t16r1s2

Note. Each multi-view pair was combined once with low, and once with high bag complexity. g = participant group (1-4), t = threat item (1-16), r = rotation (1: easy, 2: difficult), s = superposition (1: low, 2: high).

5.1.3.3 Procedure

All participants performed the test in a training classroom at Zurich Airport, where we were able to maintain standardized conditions regarding lighting, computer and monitor settings.

First, we tested the color perception of the participants using Ishihara's test of colour-blindness (2003). All participants scored 100% correct.

Within the introduction, participants were shown 16 X-ray images of guns and knives which were not used in the experiment in order for them to get an idea of what such weapons look like in X-ray images. After 4 practice trials, participants had to accomplish the 512 trials (single-view and multi-view mixed). They had to decide whether the bag presented was OK (contains no threat item) or NOT OK (contains a threat item) by clicking the respective button on the screen. Additionally, participants were asked to indicate how confident they were in their decision by clicking on a rating scale (non-visible 100 point) on the screen.

The 512 trials have been subdivided into four blocks and participants were allowed to take a short break after completing each block. Trials were randomized within each block and block order was counterbalanced across two groups of participants. Completing the experiment took about 60-90 minutes.

5.1.4 Results

We calculated the detection performance of the participants using the signal detection measure *d'*, which is derived from hit and false-alarm rates (Green & Swets, 1966). While the hit rate refers to the proportion of all images containing a prohibited item that have been judged as NOT OK, the false-alarm rate refers to the proportion of NOT OK judgments for harmless bags. The advantage of *d’* as a sensitivity measure is the fact that it is invariant when factors other than sensitivity change (Macmillan & Creelman, 2005). For calculating *d’* we used the following formula:

$$d' = z(H) - z(FA)$$

H is the hit rate, *FA* the false alarm rate and *z* refers to the z transformation.

Across all conditions, there was no significant difference in *d’* between single-view and multi-view X-ray system, $t(31) = -0.53$, $p = .30$, $d = 0.04$.

Before we examined the impact of the X-ray system on the image-based factors rotation and superposition, we analyzed the data obtained in the single-view condition to verify the existence of the image-based factors rotation and superposition as proposed by Schwaninger (2003). For this purpose, we conducted a two-way analysis of variance (ANOVA) for repeated measures using *d’* scores with the within-participant factors rotation and superposition. It revealed large main effects of rotation, $F(1, 31) = 87.17$, $p < .001$, $\eta^2 = .74$, and superposition, $F(1, 31) = 157.72$, $p < .001$, $\eta^2 = .84$. There was also a large interaction of rotation and superposition, $F(1, 31) = 31.52$, $p < .001$, $\eta^2 = .\ 50$. Figure 38 shows the effects of the image-based factors for the single-view condition. Remember that the easy rotation corresponds to a canonical view, and the difficult rotation to an 85 degree rotated view of the threat object around the vertical or horizontal axis. Superposition indicates how much the threat object is superimposed by other objects in the bag.

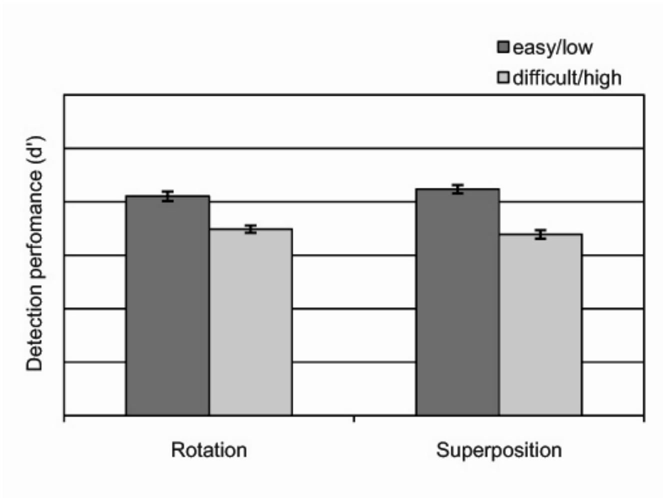

Figure 38: Effects of the image-based factors rotation and superposition for the single-view condition. Error bars represent standard errors of the mean (SEM). Note that performance values are not reported due to security reasons. However, in order to provide scientifically meaningful results, effect sizes are reported throughout the paper.

As a next step, we compared the detection performance in difficult single-view conditions with the detection performance in the corresponding difficult-easy multi-view conditions, when the first view shows the difficult single-view condition and the second view the according easy condition, i.e. the threat object was presented in a canonical rotation or, respectively, with a low superposition (this is actually the case in Figure 4). An ANOVA for repeated measures with the two within-participants factors display condition (single-view, multi-view) and image-based factor (rotation, superposition) revealed a large main effect of the display condition $F(1, 31) = 25.03$, $p < .001$, with an effect size of $\eta^2 = .45$, but not for the image-based factors $F(1, 31) = 0.24$, $p = .63$, $\eta^2 = .01$. The interaction for display condition and image-based factor was also not significant $F(1, 31) = 0.34$, $p = .56$, $\eta^2 = .01$. The pairwise comparisons revealed a significant increase of the detection performance for the multi-view condition, once for rotation, $t(31) = -3.46$, $p < .001$, with an effect size of $d = 0.27$, and once for superposition

$t(31) = -3.99$, $p < .001$, $d = 0.74$, as shown in Figure 39. According to Cohen (1988), the effect sizes are small (rotation) and medium (superposition).

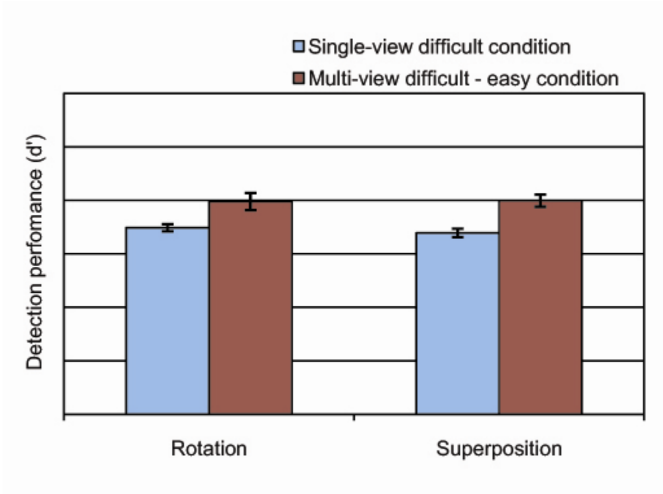

Figure 39: Effects of the display condition on the image-based factors rotation and superposition. Single-view difficult condition means a difficult rotation or a high superposition. In the difficult-easy multi-view condition, the first view was the same as in the difficult single-view condition, and the second view showed the threat object canonically rotated or with low superposition. Error bars represent standard errors of the mean (SEM). Note that performance values are not reported due to security reasons.

The increase of detection performance for the difficult-easy multi-view condition in comparison with the difficult single-view condition should also be reflected by a higher average level of confidence ratings. Therefore, we conducted an ANOVA for repeated measures using the mean value of confidence rating with the two within-participants factors display condition (single-view, multi-view) and image-based factor (rotation, superposition). We found a large main effect of the display condition, $F(1, 31) = 18.03$, $p < .001$, $\eta^2 = .37$, but not for the image-based factors $F(1, 31) = 0.78$, $p = .38$, $\eta^2 = .03$. The interaction for display condition and image-based factor was also not significant $F(1, 31) = 0.34$, $p = .56$, $\eta^2 = .01$. Pairwise comparisons disclosed a significant gain of confidence, both for rotation: $t(31) = -3.51$, p

< .001, d = 0.39, as well as for superposition, $t(31) = -3.40$, $p < .001$, d = 0.44, see also Figure 40.

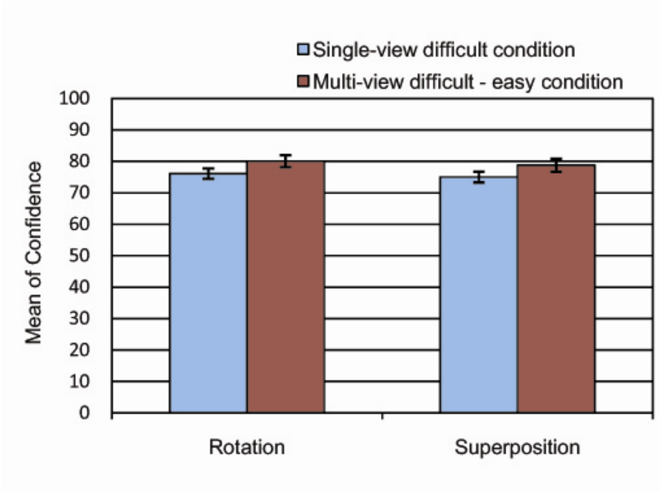

Figure 40: Effect of the display condition on the confidence ratings. Error bars represent standard errors of the mean (SEM).

Furthermore, a t-test confirmed the hypothesis that reaction times increase for the multi-view display condition in comparison with the single-view display condition, $t(31) = -12.49$, $p < .001$, d = 0.48, see also Figure 41.

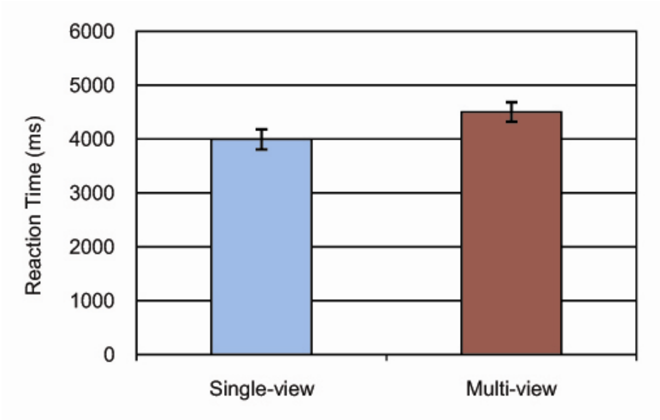

Figure 41: Effect of the display condition on the reaction time. Error bars represent standard errors of the mean (SEM).

5.1.5 Discussion

In this study, we conducted an experiment to investigate the impact of multiple views, provided by multi-view X-ray systems, on the threat object detection performance of novices. An analysis of the data pooled across all conditions revealed that there are no significant differences between the detection performance measured in *d'* for single-view and multi-view X-ray systems.

We hypothesized that the use of multi-view X-ray techniques might help especially in difficult conditions, if the threat object contained in a bag is in a difficult rotation or highly superimposed by other objects. In order to confirm this hypothesis, we first had to analyze the single-view data, to see whether we were able to replicate the effects of rotation and superposition as proposed by Schwaninger (2003). There were large main effects of rotation and superposition, and also an interaction of the two image-based factors. These results show that detection performance decreases for difficult single-view conditions.

According to our hypothesis, the threat object detection performance for the difficult conditions described above should increase, if the screening person is supported by a second view showing the threat object in an easy rotation or with a low superposition. The results of our experiment confirm this hypothesis. Moreover, the increase of detection performance with the multi-view X-ray system for difficult conditions is also reflected by a significant increase of the confidence indicated by ratings.

A probable explanation for the fact that detection performance does not increase in general for the multi-view X-ray system, but only for difficult-easy multi-view conditions might be that only in this condition the information contained in the second view is novel and helpful. In contrast, two difficult or two easy views are redundant.

Comparing the effect sizes indicates that the X-ray system has mainly an impact on superposition with a medium effect size. Whereas X-ray screening training has large effects on the detection of difficult rotated objects (Koller et al., 2008; Michel et al., 2007), Schwaninger et al. (in press) found no interaction of training and superposition. Even though our experiment revealed an effect of the X-ray system on rotation, the key benefit of multi-view X-ray systems rather seems to be the support of airport's security screeners to cope with the challenges imposed by superposition.

Furthermore, the hypothesis that reaction times increase in the multi-view condition, is confirmed, although it is only a small effect. The prolonged reaction times might be explained by the additional time needed for the visual search through two images instead of one.

In this study, we demonstrated that multi-view X-ray systems can support airport's security screeners in challenging situations, i.e. when objects in a passenger bag are difficult to recognize due to rotation or superposition. There is also a negative side-effect resulting from the multiple views: the prolonged reaction times. Previous research has shown that reaction times decrease after a certain amount of X-ray screening training (Michel et al., 2007; Schwaninger et al.,, 2007a).

Therefore, an adaptive computer-based training for multi-view X-ray systems which provides a realistic learning environment might be a way to tap the full potential of this novel technology.

5.1.6 Acknowledgement

We are very thankful to Zurich State Police, Airport Division for their help in creating the stimuli and the good collaboration for conducting the study. This research was financially supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403).

5.1.7 References

Brady, M. J. Kersten, D. (2003). Bootstrapped learning of novel objects. *Journal of Vision*, 3(6), pp. 413 – 422.

Bülthoff, H. H., Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 89, pp. 60 – 64.

Cohen, J. (1988). *Statistical Power Analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates, Hillsdale, NJ.

Graf, M., Schwaninger, A., Wallraven, C., Bülthoff, H. H. (2002). Psychophysical results from experiments on recognition & categorization. *Information Society Technologies (IST) programme, Cognitive Vision Systems – CogVis (IST-2000-29375)*.

Green, D. M., Swets, J. A. (1966). *Signal detection theory and psychophysics.* Wiley, New York.

Hardmeier, D., Hofer, F., Schwaninger, A. (2005). The X-ray object recognition test (X-ray ORT) – a reliable and valid instrument for measuring visual abilities needed in X-ray screening. *IEEE ICCST Proceedings*, 39, pp. 189 – 192.

Ishihara, S. (2003). *Ishihara's tests for colour-blindness*. Kanehara and Co, Tokyo.

Koller, S. M., Hardmeier, D., Michel, S., Schwaninger, A. (2007). Investigating training and transfer effects resulting from recurrent CBT of X-ray image interpretation. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society*, Cognitive Science Society, Austin, TX.

Koller, S. M., Hardmeier, D., Michel, S., Schwaninger, A. (2008). Investigating training, transfer, and viewpoint effects resulting from recurrent CBT of X-ray image interpretation. *Journal of Transportation Security*, 1(2), pp. 81 – 106.

Kourtzi, Z., Betts, L. R., Sarkheil, P., Welchman, A. E. (2005). Distributed neural plasticity for shape learning in the human visual cortex. *PLoS Biology*, 3(7), pp. 1317 – 1327.

Kovacs, I., Kozma, P., Feher, A., Benedek, G. (1999). Late maturation of visual spatial integration in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 96(21), pp. 12204 – 12209.

Li, W., Gilbert, C. D. (2002). Global contour saliency and local colinear interactions. *Journal of Neurophysiology*, 88(5), pp. 2846 – 2856.

Macmillan, N. A., Creelman, C. D. (2005). *Detection theory. A user's guide* (2nd ed.). Lawrence Erlbaum Associates, New York.

McCarley, J. S., Kramer, A. F., Wickens, C. D. (2004). Vidoni, E. D.; Boot, W. R.: Visual skills in airport-security screening. *Psychological Science*, 15(5), pp. 302 – 306.

Michel, S., de Ruiter, C. J., Hogervorst, M. A., Koller, S. M., Moerland, R., Schwaninger, A. (2007). Computer-based training increases efficiency in X-ray image interpretation by aviation security screeners. *41st Annual IEEE International Carnahan Conference on Security Technology*, pp. 201 – 206.

Palmer, S. E., Rosch, E., Chase, P. (1981). Canonical perspective and the perception of objects. In I. Long & A. Baddeley (Eds.), *Attention and Performance IX.* Erlbaum, Hillsdale, N.J.

Schwaninger, A. (2003). Evaluation and selection of airport security screeners. *AIRPORT*, 02/2003, pp. 14 – 15.

Schwaninger; A., Bolfing, A., Halbherr, T., Helman, S., Belyavin, A., Hay, L. (in press). The impact of image based factors and training on threat detection performance.

Schwaninger, A., Hardmeier, D., Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. *IEEE ICCST Proceedings*, 38, pp. 258 – 264.

Schwaninger, A., Hofer, F., Wetter, O. E. (2007a). Adaptive computer-based training increases on the job performance of X-ray screeners. *41st Annual IEEE International Carnahan Conference on Security Technology*, pp. 117 – 124.

Schwaninger, A., Michel, S., Bolfing, A. (2007b). A statistical approach for image difficulty estimation in X-ray screening using image measurements. *Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization*, ACM Press, New York, USA, pp. 123 – 130.

Tarr, M. J., Bülthoff, H. H. (1995). Is human object recognition better described by geon structural description or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance*, 21, pp. 1494 – 1505.

Tarr, M. J., Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. In M. J. Tarr & H. H. Bülthoff (Eds.), *Object recognition in man, monkey and machine*, MIT Press, Cambridge, MA.

Yi, D.-J., Olson, I. R., Chun, M. M. (2006). Shape-specific perceptual learning in a figure-ground segregation task. *Vision Research*, 46(6-7), pp. 914 – 924.

6 XRT Levels and Training Duration Recommendations

6.1 Relationship between level of detection performance and amount of recurrent computer-based training

6.1.1 Abstract

The usefulness of X-Ray Tutor (XRT) as an individually adaptive computer-based training system for aviation security screening officers to increase X-ray image interpretation competency has been proven in several studies. Many airports require their screening officers to conduct weekly recurrent training in order to enhance their capabilities and competencies. During training, X-ray images of passenger bags similar to how they appear at the security checkpoint have to be judged regarding the dangerousness of their content. Screening officers have to detect threat objects within the bags and discriminate harmless objects from threat objects. The advantage of XRT is its level-based construction. The training begins at the lowest difficulty level and increases in difficulty level with achieved performance for each screening officer individually. The aim of this article is to create a guideline and norm regarding training requirements in order to ensure that screening officers maximize their training benefit. Based on the X-Ray Tutor level algorithm we can express minimum requirements for screening officers in terms of image difficulty levels. For example, the minimum level to be achieved in X-Ray Tutor after one year of recurrent training was determined based on normative data assuming that screeners take about 1-2 training sessions of 20 minutes per week. Alternatively, it can also be expressed as requirement to reliably detect threat items depicted in difficult views with high superposition and medium bag complexity.

Recommendations for standard reference levels and detailed background information on the X-Ray Tutor level are given.

6.1.2 Introduction

The high potential for wide-scale fatalities on aircrafts shows the importance of scientific research in the field of airport security. Scientific research has gained high priority in the last years. Some X-ray machines provide automated detection of threat objects. Because of high false alarm rates of such technologies, the focus in X-ray image interpretation lies on the human operator who always makes the last decision if a passenger's bag is harmless or if it contains a threat object. One of the most important aspects therefore is that airport security screeners get individually adaptive training on X-ray image interpretation to enhance their knowledge about threat objects.

The X-Ray Tutor training system (XRT) was developed by the Visual Cognition Research Group (VICOREG) at the University of Zurich as a scientifically based and well-proven training program to enhance X-ray image interpretation competency very effectively (Schwaninger, 2004; Koller, Hardmeier, Michel, Schwaninger, 2007). The training program displays X-ray images of passenger bags, where screeners have to visually inspect the images and search them for threat objects like they do at the security checkpoint. XRT is based on findings about how the human brain processes visual information in order to recognise objects in different views, when superimposed by other objects, and depending on bag complexity. One core advantage of XRT is that it is individually adaptive and level-based. In other words, when screeners train with X-Ray Tutor, they reach higher difficulty levels based on their individual detection performances. The question arises which level is recommended to be reached after a specific amount of training time to guarantee a reliable and effective recognition of forbidden objects at the security checkpoint. Therefore, one important aspect is that a standard for detection performance improvement is needed, which is the main goal of this article.

Two approaches should be pursued:

1) A standard should be defined by taking into account view difficulty, superposition, and bag complexity through XRT levels.

2) The second approach is based on the data of several airports. In the course of a project, aviation security screeners at these airports conducted recurrent computer-based training with XRT during at least one year. We consider proposing a standard regarding which level in XRT screeners must have reached if they used the system for 12 months for 20 minutes per week on average.

6.1.3 Image-based and knowledge-based factors

Humans are adept at detecting different objects without any problems in a very short time as far as conditions are favourable. As soon as the conditions become unfavourable, detection performance can decrease. For example, if an object is superimposed by other objects, its shape is hard to separate (e.g., figure-ground segregation) and therefore recognition can become difficult. In the same manner, recognition of rotated objects can be difficult when objects are seen from an unusual perspective. An object can only be recognised when this particular object or a similar looking one has been seen before and been stored in visual memory. However, if there is a large difference in the appearance (e.g., because it is seen from a different angle), it cannot be recognised well anymore (always under the assumption that it has never been seen before). An object has to be seen from different angles in order to store all these views in visual memory. Then, the object can be recognised no matter which angle it is seen in. Studies regarding object recognition indicate that for most objects six views are sufficient to capture the qualitative differences resulting from viewpoint changes because the human brain is able to interpolate between stored views (Schwaninger, 2004). Additionally, if a scene (or a bag for instance) is very complex, problems in recognizing individual objects can occur. Too much information distracts attention and impedes detection and recognition of objects. These limitations that derive from the image itself are defined as image-based factors. Schwaninger, Hardmeier and Hofer (2005) defined three image-based factors as important for X-ray image interpretation (see Figure 42): It is harder to detect an object in a rotated view compared to the upright view (effect of viewpoint). The superposition of an object by others can impair the detection performance as well (effect of superposition). A bag containing different objects and the type and number of these objects can distract visual attention and therefore also impair the detection performance (effect of bag complexity). These image-based factors (view, superposition, and bag complexity) should be taken into account in an individually adaptive training for aviation security screeners.

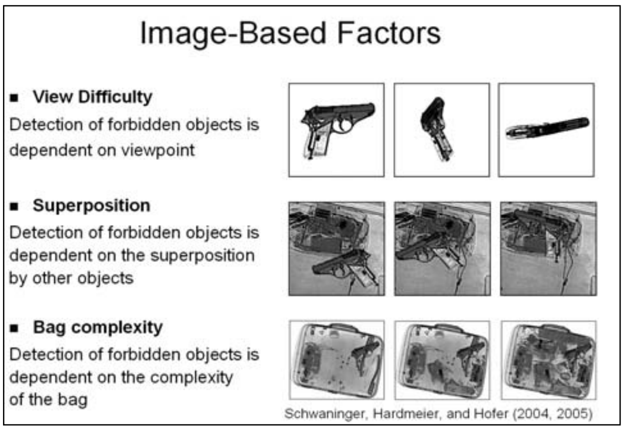

Figure 42: Explanation of image-based factors

In other words, an individually adaptive training system should increase the difficulty of the training material (i.e., X-ray images of passenger bags, some of which contain threat objects) regarding these image-based factors by means of the individual performances. For example, people are first trained with objects in easy rotations. If a certain level of detection performance in reached, the level is increased and more difficult views of threat objects are displayed. In higher levels, threat objects are more superimposed by other objects. Finally bag complexity is increased; individually adapted to the performance and difficulties a screener has in coping with each of these image-based factors. If, for example, a screener has problems coping with rotated threat objects but not when threat objects are superimposed by other objects, then the screener gets augmented training on different viewpoints of threat objects. The training proceeds in this way until the screener is able to detect rotated threat objects reliably and the screener succeeded in overcoming the difficulties in that detection performance increases. In addition, not only image-based factors are trained but also knowledge-based factors. Knowledge-based factors refer to the knowledge about which objects are prohibited and how they look like in X-ray images of passenger bags. To improve the visual knowledge, a large and representative image library is required.

6.1.4 X-Ray Tutor as an individually adaptive training system

X-Ray Tutor (XRT) is a scientifically based training program in which screeners have to decide if an X-ray image of a passenger bag is harmless or not (see Figure 43). The training is individually adaptive, that is to say, it automatically adapts to the performance of individual airport security screeners considering the difficulty of the images. X-Ray Tutor automatically combines images of fictional threat items with X-ray images of passenger bags. This is performed by an individually adaptive algorithm. X-Ray Tutor contains a large image library of threat objects that are X-rayed from different standardized views. Most of the objects can be depicted from

Figure 43: Screenshot of the XRT CBS 2.0 training system during training. At the bottom-right feedback is provided after each response. If a bag contains a prohibited item, an information window can be displayed (see bottom left of the screen).

up to 72 different viewpoints, which allows training screeners to detect threat objects independent of viewpoint. Additionally, threat objects from different threat categories (e.g., guns, knives, IEDs, other prohibited items) are integrated in XRT to make sure that a screener will be able to detect a large number of different threats. This image library was built in close collaboration with experts of Zurich State Police Airport Division and other organizations, and it is being extended continuously. The large image library is used in the individually adaptive training system in a way that objects which are poorly recognized are presented more often to a screener. How superimposed such a threat object is or the complexity of a bag that threat objects is presented in is depending on the difficulty level (see Table 18 for an overview).

How fast a screener reaches a higher level in XRT is dependant on the number of threat objects in the image library and on the screener's performance. X-Ray Tutor is available in two different versions. XRT standard edition contains 100 threat objects in up to 72 views each. XRT professional edition contains 400 threat objects also in up to 72 views each. In this article, the results are based on training with XRT standard edition.

Table 18: Difficulty levels in X-Ray Tutor

Level	Viewpoint	Super-position	Bag Complexity
1	Easy	Low	Low
2	Difficult	Low	Low
3	Easy	High	Low
4	Difficult	High	Low
5	Easy	Low	Middle
6	Difficult	Low	Middle
7	Easy	High	Middle
8	Difficult	High	Middle
9	Easy	Low	High
10	Difficult	Low	High
11	Easy	High	High
12	Difficult	High	High

During training with X-Ray Tutor, X-ray images of bags are presented on the screen for 15 seconds (standard setting). Screeners have to decide whether the bag is OK (i.e. it contains no threat object) or whether it is NOT OK (i.e. it contains a threat object). After each response, feedback is provided informing the screener whether his/her response was correct. If the bag contained a threat object the screener can view detailed information and a real image of the threat object (see Figure 2). For further information on X-Ray Tutor see Schwaninger, 2004.

6.1.5 Previous studies supporting the effectiveness of X-Ray Tutor

Different studies could show that there are large effects of viewpoint, superposition, and bag complexity due to training with XRT (Schwaniger et al., 2005; Hardmeier et al., 2006; Hofer, Hardmeier and Schwaninger, 2006). Supporting these findings, a recent study by Michel, de Ruiter, Hogervorst, Koller, Moerland and Schwaninger (2007) shows that there is an increase of detection performance not only for objects in easy but also for objects seen in difficult views. A study by Koller et al. (2007) shows that there is a very high increase of detection performance with the individually adaptive training system XRT compared to a training system which is not individually adaptive and uses a smaller image library (see Figure 44). These results are based on two measurements with the same test (X-Ray CAT, see Koller and Schwaninger, 2006) with 6 months of X-ray Tutor training between these measurements.

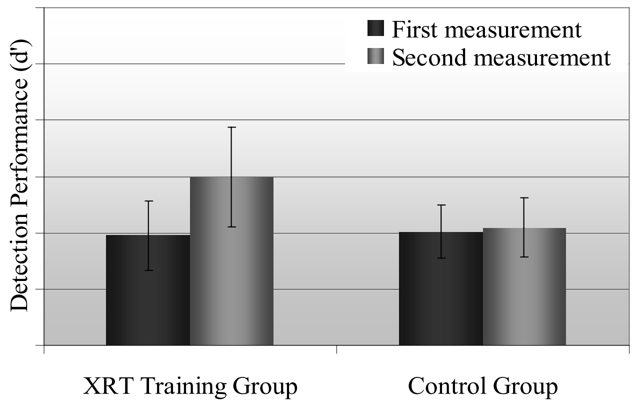

Figure 44: Detection performance with standard deviations for the XRT training group vs. the control

Another important finding is that not only trained threat objects can be recognized better due to training. It is also possible to increase recognition of visually similar objects (transfer effect). A study conducted recently by Koller et al. (2007) supports this finding. Therefore, during training the knowledge about how X-ray images of threat objects look like transfers to the detection of other similar looking objects. These findings are very important considering that not each threat object existing in the world has to be included in training. It is necessary to include objects from various categories and to have a representative threat image library. All in all the effectiveness of XRT is well proved in different studies.

6.1.6 Methods

We will mainly present the progress of the screeners' training program regarding their difficulty level increase compared to the hours of training. XRT has been operational at different airports for the same duration.

Screeners were obligated to use XRT for 20 minutes per week for at least 12 months. For example, if a screener has used the training system once a week for 20 minutes during 12 months, he has used the system for a total of 16 hours.

Table 19: Number of screeners and training hours for each airport

Airport	No. of Screeners	Training in months	Training hours (average)	Training hours (stdev.)
Airport 1	26	12	18.96	10.65
Airport 2	83	12	30.55	6.83
Airport 3	202	12	7.05	8.40
Airport 4	70	12	16.33	11.88

In total, data of 381 aviation security officers from four airports is included in this report. They trained for 14.69 hours on average with a standard deviation of about 12.97 hours during 12 months. For an overview of each of the four airports see Table 19. In this article the results are based on XRT standard edition which contains 100 threat objects in up to 72 views.

6.1.7 Recommendation for detection performance improvement

For the recommendations, two approaches should be pursued:

1) A standard should be defined by taking into account view difficulty, superposition, and bag complexity through XRT levels (see Table 1).

2) A standard should be defined by taking into account training duration and level progress.

6.1.7.1 Recommendation based on object recognition theories

A standard should be defined by taking into account view difficulty, superposition, and bag complexity through XRT levels. This minimum standard is based on object recognition theories which imply that an object can be recognised best when it has already been seen from different viewpoints. Additionally, it is also important that the forbidden objects are to some extent superimposed by other objects in the training images to make sure that objects can be well recognised independent of the position inside the bag. Finally, it is harder to find an object in a close-packed bag. Therefore, screeners should be able to recognise a threat object even if the bag is fully packed with many different harmless objects.

X-Ray Tutor contains 12 levels regarding these three image-based factors (see Chapter III). In the first level, easy viewpoints of threat items with low superposition and low bag complexity are presented. The second level contains difficult viewpoints with low superposition and low bag complexity. In level three the superposition is increased whereas the viewpoint of threat items is easy and bag complexity is low. See Table 1 for the combination of all 12 levels in X-Ray Tutor according viewpoint, superposition, and bag complexity.

Regarding these image-based factors a screener should reach level 6 after one year of training to fulfil the theoretical assumptions from object recognition theories. In level 6, a screener has seen the threat objects from easy and difficulty viewpoints. Furthermore, threat objects depicted in easy as well as in difficult orientations are presented with low and high superposition, respectively. In addition, low and middle bag complexity levels have also been seen when people have reached level 6. Therefore, screeners are exposed to almost all different image-based factors except high bag complexity. This recommendation guarantees a reliable recognition for threat objects in almost all cases at the security checkpoint. Only very high complexities are not well trained so far. When considering a bag too complex, screeners at the security checkpoint are obligated to search such a bag manually.

6.1.7.2 Recommendation based on data of airports which have trained for 12 months

A very high correlation between the amount of training hours and the XRT level achievement was found with *r* = .93.

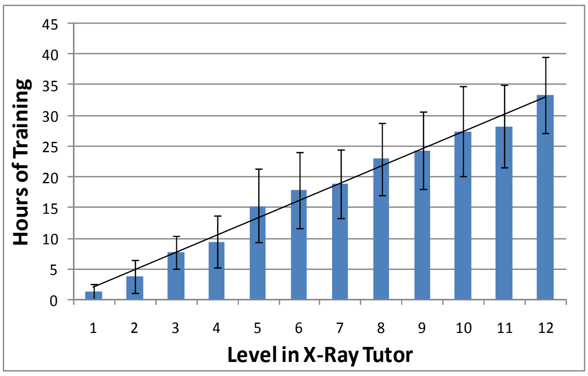

Figure 45: Average of training hours per level

Figure 45 shows the XRT level increase as a function of training duration (means and standard deviations) from real data of 381 screeners from four different airports. Note that the variation among screeners is large, with regards to how long it takes to reach a certain training level. For example, to reach level 2, 4 hours of training are needed in average. The standard deviation represents a substantial difference between people in making progress in levels.

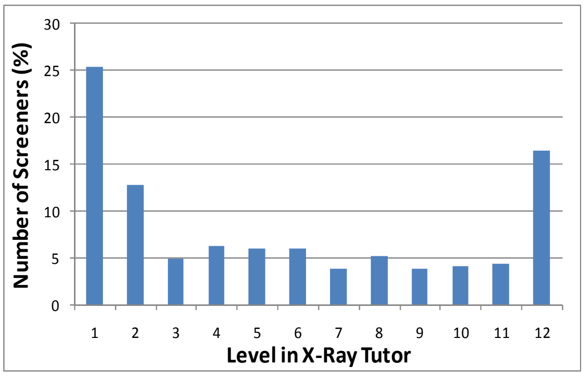

Figure 46: Number of screeners (%) per level

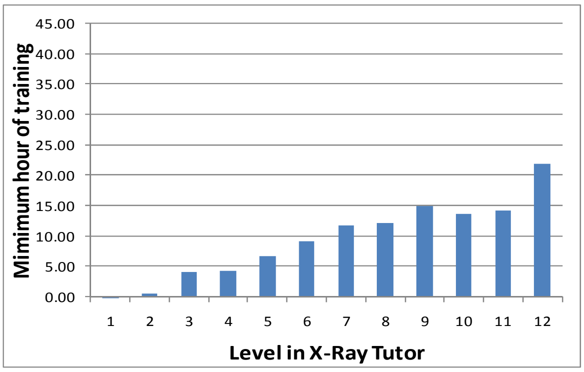

Figure 47: Minimum amount of training hours to reach a specific level in XRT.

In Figure 46, the percentage of screeners in a specific level is shown. Most screeners (25.46%) are in level 1, but none of them trained the requested amount of time of around 20 minutes per week due to different reasons (e.g., job fluctuation). Screeners in level 1 trained for about 1.37

hours within 12 months. Screeners who used the system as requested (i.e. at least 20 minutes per week during 12 months) reached level 6 after one year of training.

To conclude, based on these analyses from real data the recommendations for screeners is to reach level 6 after one year of training when XRT is used for about 20 minutes per week (in total 16 hours of training). 61.68% of all screeners in this report fulfilled this requirement and reached level 6 or higher after 12 months of training.

In addition, Figure 47 shows the minimum amount of training hours needed to reach a specific level based on the same actual data from 381 screeners. For example, the minimum amount of training time required to reach level 6 is 9.22 hours. To reach level 12 the best screener needed 21.88 hours of training.

During a training session of about 20 minutes, a screener sees about 157 X-ray images on average. In Figure 48, the numbers of images that have been seen during training are shown for each level in XRT. A very large correlation between XRT level achievement and number of images that have been seen during training was found with r = .96. The same large correlation between training duration and number of images was found with r = .96. In Figure 48 it can be seen that screeners have seen about 8'094 images on average until they reach level 6 and more than 16'606 X-ray images on average to reach level 12. With such a large number of images screeners are well

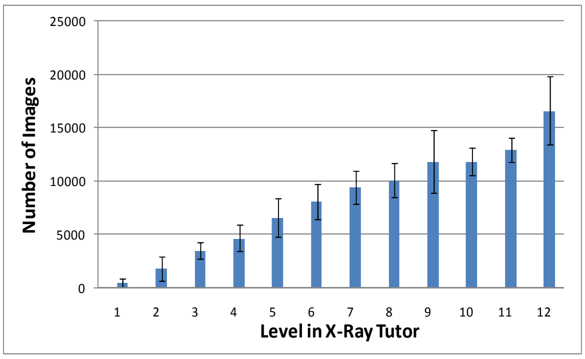

Figure 48: Number of X-ray images per level

exposed and familiarized to thousands of different X-ray images of passenger bags.

Finally, if a screener does not reach level 6 after 12 months of training, it is important that the screener is using the training system more regularly and more frequently. When screeners have reached level 12 it is important to update the image library with new threat objects for further increasing the knowledge about threat objects.

6.1.8 Conclusions

Because the human operator always has the last decision at the security checkpoint, it is important to point out that training is one of the core aspects in aviation security to guarantee a reliable and effective recognition of threat objects in passenger bags. Without training, a screener is not able to detect threat objects reliably, especially because some threat objects (e.g., IEDs) are seen very rarely at the security checkpoint. A training system should be individually adaptive to train screeners optimally based on their performances. The effectiveness of XRT has been shown in many different studies (Hardmeier et al., 2006; Michel et al., 2007; Schwaninger and Hofer, 2004; Schwaninger, Hofer and Wetter, 2007; Koller, Hardmeier, Michel and Schwaninger, 2008). For example, the study by Koller et al. (2007) showed very clear benefits for XRT. In this study, a control group trains with a non-individually adaptive training system and is compared to a group which used an individually adaptive training system (X-Ray Tutor) for the same period. Large training effects could only be found for the individually adaptive training group and not for the control group (see Figure 3). With all these findings, the importance of individually adaptive training cannot be neglected.

Another important aspect of the training system is that threat objects are presented in different difficulties like easy and difficult orientations, low and high superposition, and low and high bag complexity. This is represented in different difficulty levels implemented in XRT. The progress in these levels is dependent on the individual threat detection performances.

The aim of this article is to create a guideline and norm regarding training requirements in order to ensure that screening officers maximize their training benefits. Therefore, two recommendations based on object recognition theories and based on real XRT training data, respectively, are taken into account.

Taking the findings together, the minimum standard after 20 minutes of training per week during 12 months (is equivalent to 16 hours of training in total) is that screeners have to reach level 6 in XRT. When screeners have reached level 6, threat objects have been seen from different viewpoints, in low and high superposition, and also in medium bag complexities. This fulfils the theoretical assumptions from object recognition theories for a reliable recognition of objects.

In addition, real data shows that, on average, screeners reached level 6 when they used XRT for 12 months for 20 minutes per week. This supports the theoretical assumption and shows that it is realistically possible to fulfil these requirements.

Finally, if XRT standard edition is used to prepare screeners for the initial certification tests, it is recommended to train with XRT standard edition for about 16 hours at least or to reach level 6. Then, screeners should be well prepared for the certification test and for working at the security checkpoint.

To guarantee that the threat detection performance is stable, recurrent training with an appropriate training system and training duration (minimum 20 minutes per week) is recommended.

6.1.9 References

Hardmeier, D., Hofer, F. & Schwaninger, A. (2006). The role of recurrent CBT for increasing aviation security screeners' visual knowledge and abilities needed in x-ray screening. *Proceedings of the 4th International Aviation Security Technology Symposium, Washington, D.C., USA, November 27 – December 1,* 338-342.

Hofer, F., Hardmeier, D. & Schwaninger, A. (2006). Increasing airport security using the X-ray ORT as effective pre-employment assessment tool. *Proceedings of the 4th International Aviation Security Technology Symposium, Washington, D.C., USA, November 27 – December 1,* 303-308.

Koller, S. & Schwaninger, A. (2006). Assessing X-ray image interpretation competency of airport security screeners. Proceedings of the *2nd International Conference on Research in Air Transportation, ICRAT 2006,* Belgrade, Serbia and Montenegro, June 24-28, 399-402.

Koller, S.M., Hardmeier, D., Michel, S. & Schwaninger, A. (2007). Investigating training and transfer effects resulting from recurrent CBT of x-ray image interpretation. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society*. Austin, TX: Cognitive Science Society, 1181-1186.

Koller, S.M., Hardmeier, D., Michel, S. & Schwaninger, A. (2008). Investigating training, transfer, and viewpoint effects resulting from recurrent CBT of x-ray image interpretation. *Journal of Transportation Security, 1(2)*, 81-106.

Michel, S., de Ruiter, J.C., Hogervorst, M., Koller, S.M., Moerland, R. & Schwaninger, A. (2007). Computer-based training increases efficiency in x-ray image interpretation by aviation security screeners". *Proceedings of the 41st Carnahan Conference on Security Technology,* Ottawa, October 8-11, 2007.

Schwaninger, A. & Hofer. F. (2004). Evaluation of CBT for increasing threat detection performance in X-ray screening. In: K. Morgan and M. J. Spector, *The Internet Society 2004, Advances in Learning, Commerce and Security*, Wessex: WIT Press, 147-156

Schwaninger, A. (2004). Computer based training: a powerful tool to the enhancement of human factors. *Aviation Security International*, FEB/2004, 31-36.

Schwaninger, A., Hardmeier, D. & Hofer, F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. *IEEE Aerospace and Electronic Systems, 20(6),* 29-35.

Schwaninger, A., Hofer, F. & Wetter, O.E. (2007). Adaptive computer-based training increases on the job performance of x-ray screeners. *Proceedings of the 41st Carnahan Conference on Security Technology,* Ottawa, October 8-11.

7 Danksagung

Die visuelle Wahrnehmung in einem angewandten Bereich wie der Flughafensicherheit zu untersuchen, und dabei Wege zu finden, wie die Erkennung von verbotenen Gegenständen mittels Training verbessert werden kann, ist ein ebenso relevantes wie spannendes Thema. Heutzutage ist es, wie oft in modernen Forschungsgebieten so, dass durch die Zusammenarbeit mit anderen Forschern, die komplexen Prozesse unseres kognitiven Apparates besser verstanden und dadurch wichtige Forschungsfragen beantwortet werden können. In dieser Doktorarbeit wurde ich von verschiedenen Personen unterstützt und gefördert. Die verschiedenen Experimente dieser Arbeit sind in hervorragender Kollaboration mit anderen Forschern entstanden. Bei Markus Ruh bedanke ich mich für die tolle Zusammenarbeit in Kapitel 2. Kapitel 3 habe ich in sehr guter Zusammenarbeit mit Anton Bolfing durchgeführt. Bei Saskia Koller möchte ich mich für die wertvolle Kollaboration in Kapitel 2 und Kapitel 4 bedanken. Ebenfalls bedanken möchte ich mich bei Diana Hardmeier für die tolle Unterstützung in Kapitel 4. Bei Claudia von Bastian bedanke ich mich ganz herzlich für die tolle Zusammenarbeit in Kapitel 5. Des Weiteren bedanke ich mich herzlich bei verschiedenen Kollegen und Kolleginnen für die wertvollen Diskussionen. Betonen möchte ich, dass alle wissenschaftlichen Überlegungen und Ideen zu den durchgeführten Studien, inklusive der Studiendesigns, von mir stammen. Ich wurde lediglich in der Durchführung der einzelnen Experimente von den oben erwähnten Personen unterstützt.

Meinem Betreuer, Prof. Dr. Adrian Schwaninger, möchte ich einen besonderen Dank für die hervorragende Unterstützung aussprechen.

Prof. Dr. Wolfgang Marx möchte ich ebenfalls einen besonderen Dank aussprechen. Schon während dem Studium und später während der Dissertation hat er mich wertvoll unterstützt, wofür ich ihm sehr dankbar bin.

Prof. Dr. Martin Kleinmann danke ich ganz herzlich für seine Bereitschaft, meine Dissertation als Zweitreferent zu begutachten.

Den Arbeitskollegen und Kolleginnen an der Universität Zürich danke ich für die wertvollen Diskussionen und die gute Zusammenarbeit auch bei administrativen Dingen.

Die spannenden Fragen hätten nicht ohne Kollaborationen mit verschiedenen europäischen Flughäfen beantwortet werden können. Besonders bedanken möchte ich mich bei den betreffenden Mitarbeitern und Kaderpersonen der Kantonspolizei Zürich, Flughafenpolizei.

Besonders danken möchte ich meinen Eltern für die emotionale und finanzielle Unterstützung während der ganzen Doktorandenzeit. Sie waren immer für mich da und haben mir das Studium erst ermöglicht. Ohne diese Unterstützung wäre es nicht möglich gewesen, diese Dissertation zu schreiben.

Meinen Freunden danke ich ebenfalls für die emotionale Unterstützung während der Doktorarbeit.

Last but not least danke ich allen Versuchspersonen, welche mit grosser Motivation an meinen Experimenten teilgenommen haben.

8 Curriculum Vitae

PERSONAL INFORMATION

Name	STEFAN MICHEL
Address	UNGARBÜHLSTRASSE 74, CH-8200 SCHAFFHAUSEN
Telephone	+41 78 611 47 80
E-mail	s.michel@psychologie.uzh.ch
Nationality	Switzerland
Date of birth	01.09.1977, Schaffhausen, Switzerland

WORK EXPERIENCE

• Dates (from – to)	Since 2006
• Name and address of employer	University Zurich, Binzmühlestrasse 14/22, CH-8050 Zürich
• Type of business or sector	Department of Psychology, General Psychology (Cognition), Visual Cognition Research Group
• Occupation or position held	Doctoral Student (project manager training studies)
• Dates (from – to)	Since 2006
• Name and address of employer	Max Planck Institute for Biological Cybernetics, Speemannstr. 38, D-72076 Tübingen
• Type of business or sector	Cognitive and Computational Psychophysics Department
• Occupation or position held	Doctoral Student
• Main activities and responsibilities	Manager EU project
• Dates (from – to)	Since 2006
• Name and address of employer	University Zurich
• Main activities and responsibilities	Lecturing
• Dates (from – to)	2004 – 2006
• Name and address of employer	University Zurich, Binzmühlestrasse 14/22, CH-8050 Zürich
• Type of business or sector	Department of Psychology, General Psychology (Cognition), Visual Cognition Research Group
• Occupation or position held	Research Assistant
• Main activities and responsibilities	Experimental psychology studies
• Dates (from – to)	2002 – 2004
• Name and address of employer	Applied Psychological Science Solutions (APSS GmbH)
• Occupation or position held	Research Assistant
• Main activities and responsibilities	Training development

• Dates (from – to)	2002 - 2003
• Name and address of employer	Unilever Bestfood, CH - 8240 Thayngen
• Occupation or position held	Production Assistant

EDUCATION AND TRAINING

EDUCATION:

• Dates (from – to)	1999 - 2006
• Name and type of organization providing education and training	University Zurich
• Principal subjects/occupational skills covered	Studies of psychology, child and adolescent psychiatry and neuropsychology
• Title of qualification awarded	Thesis Title (Lizentiat, equivalent to Master's Degree): "The Role of Co-Occurrence for View-Based Object Recognition"
• Level in national classification (if appropriate)	M Sc UZH

• Dates (from – to)	2003 – 2004
• Name and type of organization providing education and training	Department of Child and Adolescent Psychiatry
• Occupation or position held	Internship
• Main activities and responsibilities	ABA (Applied Behavior Analysis) Therapy for Autism

• Dates (from – to)	2003
• Name and address of employer	University of Dundee
• Occupation or position held	Internship
• Main activities and responsibilities	Inhibition of Return in Dynamic Gaze Cue Paradigms

• Dates (from – to)	1994 - 1999
• Name and type of organisation providing education and training	Kantonsschule Schaffhausen
• Title of qualification awarded	Matura Typus C

TRAINING:

• Dates (from – to)	December 2007
• Name and type of organisation providing education and training	University Zurich, Didactica
• Principal subjects/occupational skills covered	Course: Projektmanagement bei der Betreuung von Diplom- und Lizentiatsarbeiten

• Dates (from – to)	November 2007
• Name and type of organisation providing education and training	University Zurich, Informatikdienste
• Principal subjects/occupational skills covered	Course: Excel: Makroprogrammierung mit VBA

SOCIAL SKILLS AND COMPETENCES *Living and working with other people, in multicultural environments, in positions where communication is important and situations where teamwork is essential (for example culture and sports), etc.*	Team Work acquired as project manager training studies Coaching of students Communication Skills: Communication skills acquired and developed during teaching courses for students, working close to undergraduate students, guiding them in achieving theoretical and practical skills in the field of research methods, oral presentations on the results of my researches: as researcher and at international meetings and conferences Member of ice hockey team for 7 years
ORGANISATIONAL SKILLS AND COMPETENCES *Coordination and administration of people, projects and budgets; at work, in voluntary work (for example culture and sports) and at home, etc*	Organizational skills acquired as project manager training studies - Requires planning, coordination of different tasks and collaborators to meet submission deadlines. During the work at VICOREG acquiring good skills in coordinating projects, dealing with project partners and customers, leading the visual cognition project group at University of Zurich and supervise undergraduate work.
TECHNICAL SKILLS AND COMPETENCES *With computers, specific kinds of equipment, machinery, etc.*	Fully competent with most Microsoft computer programmes (MS Windows, MS Excel, MS Word, MS Powerpoint, MS Project, MS Access, MS Outlook), Adobe Photoshop, SPSS, experimental software (E-Prime) Good knowledge in testing software (algorithms and usability) Good knowledge in software- and network installation (domain server) Competent in quantitative and qualitative research methods, experimental designing, data analysis Basic knowledge in Visual Basics for Application, LAN technology
ARTISTIC SKILLS AND COMPETENCES MUSIC, WRITING, DESIGN, ETC.	5 years trumpet training 2 years electronic keyboard training Currently playing the Didgeridoo and the Djembe Amateur photographer
OTHER SKILLS AND COMPETENCES *Competences not mentioned above.*	Highly multitasking competent Work adaptively and effectively under pressure Good ability to think and work independently
DRIVING LICENCES	Car and motorbike, both acquired in Switzerland

MOTHER TONGUE	GERMAN
OTHER LANGUAGES	
	ENGLISH
• Reading skills	excellent
• Writing skills	good
• Verbal skills	good
	FRENCH
• Reading skills	basic
• Writing skills	basic
• Verbal skills	basic

ADDITIONAL INFORMATION

COLLABORATIONS

Research Institutes

- Department of Industrial Engineering, University at Buffalo: SUNY, U.S.A. (Prof. C. Drury)
- Department of Psychology, University of Dundee (Dr. Martin Fischer)
- Max Planck Institute for Biological Cybernetics, Tübingen, Germany (Prof. H.H. Bülthoff)
- TNO Prins Maurits Laboratory, The Netherlands (C.J. de Ruiter)

Government

- Belgian Civil Aviation Administration (F. Durinckx)
- European Aviation Security Training Institute (EASTI), Brussels, Belgium (F. Durinckx)
- Ministry of Interior, Germany (R. Strassmann)
- Ministry of Justice, The Netherlands (project with TNO)

Companies

- Zurich State Police, Airport Division (W. Wüest)
- Amsterdam Airport Schiphol (R. Jongejan)
- Brussels International Airport Company (C. Coutereel)
- Security Training International (W. Glassner)
- Siemens Switzerland (M. Städler)
- Smiths Heimann International (H.-J. Schöpe)
- Zurich Airport Unique (E. Schlecht)

Teaching Experience

Spring Semester 2008

• *No.*	2299
• *Lecturers*	Stefan Michel & Anton Bolfing
• *Title*	Expra (Arbeits- und Organisationspsychologie)
• *No.*	2370
• *Lecturers*	Wolfgang Marx, Daniel Hausmann-Thürig, Stefan Michel, Stefan Ryf
• *Title*	Allgemeinpsychologisches Seminar: Erleben - Entscheiden – Handeln
• *No.*	2444
• *Lecturers*	Sarah Chiller-Glaus, Diana Hardmeier, Stefan Michel, Adrian Schwaninger
• *Title*	Projektgruppe Wahrnehmung und Visuelle Kognition

Summer Semester 2007

• No.	2411
• Lecturers	Sarah Chiller-Glaus , Stefan Michel, Adrian Schwaninger
• Title	Visual Cognition Research Group

Winter Semester 2006/2007

• No.	2155
• Lecturers	Adrian Schwaninger, Stefan Michel
• Title	Methodenpropädeutikum II (Literaturkolloquium): Allgemeine Psychologie
• No.	2250
• Lecturers	Wolfgang Marx, Daniel Hausmann-Thürig, Stefan Michel, Stefan Ryf
• Title	Allgemeinpsychologisches Seminar: Intuition & Reflexion

PUBLICATIONS

Peer Reviewed Articles

Koller, S.M., Hardmeier, D., Michel, S., & Schwaninger, A. (2008). Investigating training, transfer, and viewpoint effects resulting from recurrent CBT of x-ray image interpretation. Journal of Transportation Security, 1/(2), 81-106.

Michel, S., de Ruiter, J. C., Hogervorst, Koller, S.M., M., Moerland, R., & Schwaninger, A. (2007). Computer-based training increases efficiency in x-ray image interpretation by aviation security screeners. Proceedings of the 41st Carnahan Conference on Security Technology, Ottawa, October 8-11, 2007.

Koller, S.M., Hardmeier, D., Michel, S., & Schwaninger, A. (2007). Investigating training and transfer effects resulting from recurrent CBT of x-ray image interpretation. Proceedings of the 29th Meeting of the Cognitive Science Society, 1181-1186.

Michel, S., Koller, S.M., Ruh, M., & Schwaninger, A. (2007). Do "image enhancement" functions really enhance x-ray image interpretation? Proceedings of the 29th Meeting of the Cognitive Science Society, 1301-1306.

Schwaninger, A., Michel, S., & Bolfing, A. (2007). A statistical approach for image difficulty estimation in x-ray screening using image measurements. Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization, ACM Press, New York, USA, 123-130.

Bolfing, A., Michel , S., & Schwaninger, A. (2006b). A statistical approach for automated image difficulty estimation in x-ray screening using image processing algorithms. Proceedings of the 4th International Aviation Security Technology Symposium, Washington, D.C., USA, November 27 – December 1, 2006, 384-388.

Michel, S., Koller, S., Ruh, M., & Schwaninger, A. (2006). The effect of image enhancement functions on X-ray detection performance. Proceedings of the 4th International Aviation Security Technology Symposium, Washington, D.C., USA, November 27 – December 1, 2006, 434-439.

Bolfing, A., Michel , S., & Schwaninger, A. (2006a). Assessing image difficulty in x-ray screening using image processing algorithms. Proceedings of the 2nd International Conference on Research in Air Transportation, ICRAT 2006, Belgrade, Serbia and Montenegro, June 24-28, 2006, 253-258.

Schwaninger, A., Michel , S., & Bolfing A. (2005). Towards a model for estimating image difficulty in x-ray screening. IEEE ICCST Proceedings, 39, 185-188.

Reports

Schwaninger, A., Hogervorst, M.A., Michel, S., Koller, S., & de Ruiter, C.J. (2007). Impact of X-Ray Tutor training and image quality on screener performance at Amsterdam Airport Schiphol, TNO Report.

CONFERENCE ABSTRACTS

Michel, S., Koller, S.M., Ruh, M., & Schwaninger, A. (2007). Are "Image Enhancement Functions" Really Useful To Enhance X-Ray Images? Poster at Presentation at the 10th Congress of the SGP/SSP, Zurich, Switzerland.

Hofer, F., Schwaninger, A., & Michel. S. (2003). Top-Down Effekte in der Objekterkennung. 45. Tagung experimentell arbeitender Psychologen (TEAP), Kiel, Germany.

Hofer, F., Schwaninger, A., & Michel. S. (2003). Top-Down Effects in Object Recognition. Presentation at the 8th Congress of the SGP/SSP, Bern, Switzerland.

Michel, S., Bolfing, A., Hofer, F., & Schwaninger, A. (2004). Towards a Psychophysically Plausible Model of X-Ray Threat Detection Performance. Poster presented at the 2nd LIDOKO (Lizentianden und Doktorandenkongress), Zurich.

Michel, S., Hofer, F., & Schwaninger, A. (2003). Top-Down Effekte in der Objekterkennung, Poster presented at the 1st LIDOKO (Lizentianden und Doktorandenkongress), Zurich.

Schwaninger, A., Michel., S. Hofer, F., & Bülthoff, H.H. (2003). The

Role of Co-occurence of Objects for View-Based Object Recognition. In: H.H. Bülthoff, K. Gegenfurtner, H.A. Mallot, R. Ulrich & F.A. Wichmann (Hrsg), TWK 2003, 6. Tübinger Wahrnehmungskonferenz. Kirchentellinsfurt: Knirsch.

INVITED LECTURES AND PRESENTATIONS

Michel, S. (2008). The effect of single versus multi-view display on threat detection, InterTAG Meeting, 27. – 29. May, Toulouse, France

Michel, S. (2007). How To Test Performance On The Job Using Tests Built Into CBT. Development of reference levels framework for aviation security screeners, VIA Meeting, September 26-28, Berlin.

Michel, S. (2007). How To Enhance The Quality Of Human-Machine Interaction With Regard To Image Display Technology, VIA Meeting, September 26-28, Berlin.

Michel, S., Koller, S., Ruh, M., & Schwaninger, A. (2007). Are image enhancement functions really enhancing x-ray detection performance? In Wender, K. F., Mecklenbräuker, S., Rey, G. D., Wehr, TH. (Eds.), Presentation at the 45. Tagung experimentell arbeitender Psychologen (TEAP), Trier, Germany.

Michel, S. (2007). Steigern "Image Enhancement Functions" die Erkennung von verbotenen Gegenständen in Röntgenbildern? Introduction of VICOREG Group at the lecture: Fundamentals of General Psychology 1: Perception, March 3, Zurich.

Michel, S. (2006). Assessing Image Difficulty in X-Ray Screening Using Image Processing Algorithms. 17th International Aviation Security Human Factors Technical Advisory Group Meeting (InterTAG), October 3, Zurich, Switzerland.